

MIRages: an account of music audio extractors,  
semantic description and context-awareness,  
in the three ages of MIR

Perfecto Herrera Boyer

---

TESI DOCTORAL UPF / ANY 2018

DIRECTORS DE LA TESI

Dr. Xavier Serra i Casals

Dra. Emilia Gómez Gutiérrez

DEPARTAMENT DE TECNOLOGIES DE LA INFORMACIÓ I LES  
COMUNICACIONS



**FORTITUDINE VINCIMUS**

(by endurance we conquer)

Family *motto* of the British polar explorer Ernest Shackleton



## Acknowledgements

Thanks

To my late and mournfully missed parents,  
They gave me the song, the word and the number  
They encouraged my curiosity and perseverance  
They showed me the way. And by endurance we all conquered.

To my dearest, sweetest, and cheerful Eulalia.

This thesis brought boredom, sadness and distress to her life. There will not be enough time or love from me to pay off for all that. I cannot conceive a way or a moment when she will be rewarded for what she was stolen during this long and horrid journey by my side. I wish she smiles when seeing this task successfully finished.  
By endurance, she will conquer.

It has been a long time since I phoned Xavier Serra to ask him for me to be accepted in a computer music course he was about to teach in the Phonos Foundation in Barcelona. His name sounded familiar to me from some Computer Music Journal I, by chance, browsed some months before in the library of the only public place in the pre-Olympic Barcelona where you could listen to and borrow electronic and computer music recordings. By chance, too, I had read about that course in a local classical music magazine that I had never browsed before. My guts were telling me that I could not miss it, and they were right, as usual: it changed my life for good and forever. It was 1992 and, since then, Xavier and I have shared many events, conditions, projects, meetings and journeys and I have never regretted of that: working and enjoying it at the same time is priceless, and he gave me the opportunity to achieve that. Being part of the MTG since its very beginning has been an honour and a permanent challenge. I am also grateful to him for the lessons he (on purpose or unwittingly) taught me about technical, scientific, musical and management issues. I would wish not having betrayed the trust he put on me, and I hope this thesis helps him to feel relieved about me and about his academic duties. We have shared a long intellectual journey that I would not like to end here. He deserves my earnest, genuine and permanent gratitude and admiration. What he has achieved with the MTG was unbelievable twenty-five years ago, considering we live in a country with so many rude, ignorant, unscientific and music-insensitive people.

Emilia Gómez provided the right tonality to many years of research. In our shared projects, she always held the metronome, the compass and the timetable and helped me to keep many things in order, on the right track, and on time. Her enthusiastic support, collaborative attitude and rigor are outstanding, and I could not have been luckier and, at the same time, more indebted with such a proactive and tireless long-term research partner and, in the recent years, thesis supervisor. I also owe her the idea of trying again to set up this manuscript and the final push for reaching a happy end along that path. A master on hitting the right chords, indeed.

Many colleagues have been valuable sources of knowledge, critical thought, enlightenment, fun and encouragement (different linear combinations apply here). Working with them has been a pleasure and an occasion for intellectual and personal growth. Sergi Jordà has been a long-time partner in many adventures, long before the

ages of MIR. We have shared and debated about music, art, science, lifestyles, scents, politics, hobbies... Even though our projects, skills and interests were rarely convergent during many years I am very grateful he rescued me from a personal and professional black hole for the GiantSteps project. I appreciate his sincere critical and bold perspectives, his enthusiasm and will to act, to make, to change the world. He helps me to overcome my slow, stratospheric way of thinking to go down to earth quickly, where practical problems and real users must be promptly faced. I think we make an interesting, complementary yin-yang duet (probably better when doing research than when hitting our synths and noisy devices). Improvising, wandering about, or following straight paths with him (depending on the task, mood or deadline) has been a rewarding and game-changing experience. Joan Serrà was instrumental for getting a former thesis draft that never was concluded but helped to frame the current one. He played, without any reward, the role of “bad cop” during many months and showed infinite patience with my frequent delays on our agreed deadlines. He pushed me gently to the edge, making me feel that the task of my thesis had a sense, a purpose and a possible direction (even when, afterwards, that initiative was ground to a halt). Just because of that, my debt towards him is immense. Versions of us even explored amazing territories beyond our original areas of expertise, and complex networks never were the same after that. In all those activities he provided to me excellent examples of rigor, systematicity and forward-thinking. From time to time I wish many of my supervised students would be versions of him. Oscar Celma was not a FOAF, but he was a master of many trades, a native explorer and a proficient provider of workable solutions to almost everything. Without his hacking zest and semantic awareness, nothing about contextual issues could have been addressed the same way. I hope he can keep the Pandora box tightly closed for many years. Fabien Gouyon layered the early rhythms of my research and showed me the way to be systematic, open-minded and multi-disciplinary. His vibrant intellectual pulse kept many challenges active in my research life, and I thank him the activities in which, from time to time, he has proposed me to tick along. Geoffroy Peeters was another early and essential influence. The many occasions, days and kilometres travelled together gave me the chance to learn about signal processing, methodology and music retrieval from a true visionary. I wish we would have had more time to indulge into conversing about music and life issues with a beer on the table. Dmitry Bogdanov has been a bright, friendly and loyal research partner for more than a decade, and I have always felt honoured by his kind invitations to generate ideas, to sequence projects or to master ongoing papers. In addition to tuning, filtering, amplifying and playing with research projects, we easily indulged on discussions about analogue synths and music production techniques (which sometimes was even more exciting than the original purpose of our meetings). It has always been reassuring to feel connected with him at levels transcending research but that exert an influence on it.

Even though I would like to write long sentences to express gratitude to most of my former and current collaborators, I will refrain from doing that, as I would need several pages (they are nearly a hundred!). The following list includes not only direct collaborators but also MTG colleagues with whom I have enjoyed many sparkling or boring discussions, lunch or coffee time, accepted or rejected project proposals and papers, exciting or useless committees, project meetings or informal meetings. They are (my apologies if I missed somebody): Vincent Akkermans, Xavier Amatriain, Josep Lluís Arcos, Pau Arumí, Jean Julien Aucouturier, Thomas Aussenac, Eduard Aylon, Giuseppe Bandiera, Eloi Batlle, Juan Pablo Bello, Martin Blech, Jordi Bonada, Juanjo Bosch, Paul

Brossier, Pedro Cano, Rafael Caro, Álvaro Corral, Manuel Davy, Marteen deBoer, Amaury Dehamel, Sue Denham, Simon Dixon, Shlomo Dubnov, Ángel Faraldo, Frederic Font, Ferdinand Fuhrmann, Chen Gang, David García, José-Pedro García, David García, Daniel Gómez, Mastaka Goto, Marteen Grachten, Jens Grivolla, Sankalp Gulati, Martín Haro, Amaury Hazan, Martin Hermant, Piotr Holonowicz, Henkjan Honing, Leonidas Ioannidis, Jordi Janer, Stefan Kersten, Peter Knees, Stefan Koelsch, Markus Koppenberger, Cyril Laurier, Sylvain Le Groux, Ramón López de Mántaras, Alex Loscos, Esteban Maestre, Marco Marchini, Agustín Martorell, Ricard Marxer, Jordi Massaguer, Jaume Massip, Óscar Mayor, Owen Meyers, Felipe Navarro, Javier Nistal, Waldo Nogueira, Cárthach Ó Nuanáin, Bee-Suan Ong, François Pachtet, Elias Pampalk, Panos Papiotis, Alfonso Pérez, Gilles Petterschmitt, António Sá Pinto, Alaister Porter, Hendrik Purwins, Miquel Ramirez, Rafael Ramírez, Zuriñe Resa, Julien Ricard, Martín Rocamora, Gerard Roma, Carlos Román, Justin Salamon, Inês Salselas, Mark Sandler, Vegard Sandvold, Álvaro Sarasúa, Markus Schedl, Mattia Schirosa, Leigh Smith, Mohamed Sordo, Sebastian Streich, Sergio Toral, Julián Urbano, Hughes Vinet, Nicolas Wack, Gerhard Widmer, Istvan Winkler, Anna Xambó, Yi-Hsuan Yang, Alex Yeterian, Massimiliano Zanin and José Zapata. In addition to their nice work, they contributed to create the friendly and inspiring intellectual environment in which my work matured. I hope to have given to them, at least, as much as I got from them.

Many students in the Sound and Music Computing Master and in the Sonology degree (and colleagues in the Escola Superior de Música de Catalunya) should also be credited for challenging my views, for posing the right questions that raised my awareness of what I didn't know (and I don't know yet) or for encouraging me to keep on the track (or the opposite way). I also felt lucky of having at hand the wisdom and extensive experience of Gabriel Brncic and Andrés Lewin, who provided, in many occasions, music, technology, and personal life lessons, in addition to encouragement to pursue my career and finish this thesis. Administrative and research support people as Cristina Garrido Lydia García, Joana Clotet, José Lozano, Sonia Espí, Ramón, Loureiro, Salvador Gurrera, Rosa Borràs, Alba Rosado and Carlos Atance, made the MTG and the UPF an efficient, friendly and well-organized place (as I have been told by many people, *like no other working environment in the world* -though I miss a lot our former HQ in Ocata-Estació de França!-).

Insights and experiences far from the academic field were possible from my years of professional practice in sound recording, reinforcement and mixing. I thank Pere Aguilar, my former business partner in those activities, for the lessons he taught me, and for the studio opportunities he immersed me into. And I should not forget Juanjo Tortosa, who gave me the very first learning chances in a professional studio. Those years, activities and frustrations planted several seeds for my posterior interests and drives in music technology research.

I started my research career in a very different place (University of Barcelona's Departament de Psicologia Bàsica) and field (Cognitive Psychology), as one of the lucky receivers of the first scholarships that the Generalitat de Catalunya provided to PhD students (that was in 1989!). I guess I supplied the Catalan Government with the first evidence that the scholarship system was not failure-proof as I quitted in my 3rd year without finishing my thesis on *implicit learning*. Personally, I never experienced it as a

failure but as a kind of liberation from an unhappy (and maybe easy) track that was not leading me towards the (physical, academic and mental) place I wanted to be. It also brought me the opportunity to follow my then-confirmed musical/technological vocation. During that period in the UB, I learnt many research skills, ethics and attitudes from two key figures with which I am in deep debt too: my former supervisor Elisabet Tubau, and my informal mentor Núria Sebastian. Most of what I learnt about research with them or from them could, fortunately, be adapted or re-used in my posterior research activity so, in the end, they should be happy for their time and efforts not being wasted at all after my disbanding. During that period (and also in more recent years), I also learnt a lot about how not doing a PhD thesis, and maybe I will write a book on that in the future, given the experience I have been amassing in so many years of procrastination and diletantism.

I cannot finish this section without a remembrance of some influential teachers that I enjoyed in my High-school years. Agustín Montori, Javier Peidró and Félix Allí showed me the way to learn and love music, mathematics and writing, respectively. I never then imagined these disciplines would provide the building blocks of my professional career. Therefore, I owe gratitude to them for channelling to me the fun in calculus, the beauty in geometry, the light in the right words, the power of well-crafted paragraphs, the harmony in the discord and in the noise, the musicality in everyday sounds. They all provided priceless lessons to appreciate the differences between dogmatic and open attitudes, between indolence and curiosity.

*Takk takk* to my Ph. Doctor friends Álex, Óscar and Vicenç, who were always supportive and/or critical (according to the moment) in all this long and winding road, as in many other aspects in my life. *Með suð í eyrum við spilum endalaust*. Jesús and Ricard were also supportive whenever this topic was on the after-lunch conversations (though conversations dealt mostly on more interesting matters, fortunately). Thanks, and long life to all of them, for me to keep on counting with so many different perspectives on science, art, music, life and ways of having fun and growing as a person.

The only living being mentioned here for whom music did not meant anything was Benito, our eternally missed lady cat. She was cat-egorically making my life sparkling. She was a teacher of essential lessons, the *medium* that took me outside of my own self, the provider of purr-spective and a cat-a-list of emotions. She took naps in and sat on the cardboard box of the computer in which early versions of this meow-nuscript were written; sometimes she dared even typing some lines for me, and she enjoyed many referenced papers (in the times when papers were truly paper-made) more than I did. Her endurance conquered my heart.

I cannot see my work detached from the rest of my life or just becoming the product of a moment. As I have been resonating with the input of all these beings and experiences I am thankful of, I wanted to say so. I hope the reader is pleased with the hum, buzz and drone that is coming out and ensuing.



## Abstract

This thesis reports on research carried out and published during the last twenty years on different problems of Music Information Retrieval (MIR). We organize the text as a personal account and critical reflection along four hypothesized ages that have shaped the evolution of MIR. In *the age of feature extractors*, we present work on features to describe sounds and music, especially timbre and tonal aspects. In *the age of semantic descriptors* work on describing music with high-level concepts, such as mood, instruments, similarities, cover versions or genres, usually inferred with machine learning from annotated collections is reported. In *the age of context-aware systems* we report on user models for recommendation and for avatar generation, in addition to factors that influence music listening decisions. We finally discuss the possibility of a more recent *age of creative systems* where MIR features, classifiers, models and evaluation methodologies aid to enhance or expand music creation.

## Resum

Aquesta tesi informa sobre recerca realitzada i publicada durant els últims vint anys en diferents problemes de Recuperació d'Informació Musical (MIR). Organitzem el text com a visió personal i reflexió crítica i utilitzant quatre hipotètiques edats que han configurat l'evolució del MIR. A l'*edat dels extractors de característiques*, presentem treballs sobre trets per a descriure sons i música, especialment timbre i aspectes tonals. A l'*edat dels descriptors semàntics* es treballa en la descripció de música amb conceptes d'alt nivell, com l'estat d'ànim, els instruments, les similituds, les versions musicals o els gèneres, generalment deduïts amb l'aprenentatge automàtic a partir de col·leccions anotades. En l'*era dels sistemes sensibles al context*, informem sobre models d'usuaris amb l'objectiu de fer recomanacions musicals i generació d'avatars, a més de factors que influeixen en les decisions d'escoltar música. S'esmenta, finalment, una possible i més recent *edat dels sistemes creatius* on els descriptors, classificadors, models i metodologies d'avaluació de MIR ajuden a potenciar o ampliar la creació musical.



# Table of contents

<b>Acknowledgements</b> .....	v
<b>Abstract</b> .....	ix
<b>Resum</b> .....	ix
<b>Table of contents</b> .....	xi
<b>List of figures</b> .....	xiii
1. THE THREE AGES OF MIR.....	1
2. THE AGE OF FEATURE EXTRACTORS .....	5
2.1. Introduction .....	5
2.2. Papers included in this chapter .....	8
2.3. Contributions in the selected papers.....	8
3. THE AGE OF SEMANTIC DESCRIPTORS.....	75
3.1. Introduction .....	75
3.2. Papers included in this chapter .....	77
3.3. Contributions in the selected papers.....	78
4. THE AGE OF CONTEXT-AWARENESS .....	163
4.1. Introduction .....	163
4.2. Papers included in this chapter .....	165
4.3. Contributions in the selected papers.....	166
5. THE AGE OF CREATIVE SYSTEMS? .....	195
5.1. Introduction .....	195
5.2. Papers included in this chapter .....	198
5.3. Contributions in the selected papers.....	198
6. CONCLUDING THOUGHTS.....	219
REFERENCES .....	223



## List of figures

Fig. 1.1 Timeline with articles included in the age of feature extractors .....	Pg. 3
Fig. 1.2. Timeline with articles included in the age of semantic descriptors.....	3
Fig. 1.3. Timeline with articles included in the age of context-aware systems...	4



# 1. THE THREE AGES OF MIR

Music Information Retrieval or, as some of us prefer, Music Information Research, is a multidisciplinary field shaped during the last two decades and which is still struggling to find its place in the world. According to some authors, “MIR is a multidisciplinary research endeavour that strives to develop innovative content-based searching schemes, novel interfaces, and evolving networked delivery mechanisms in an effort to make the world’s vast store of music accessible to all” (Downie, 2004) and “is concerned with the extraction, analysis, and usage of information about any kind of music entity (for example, a song or a music artist) on any representation level (for example, audio signal, symbolic MIDI representation of a piece of music, or name of a music artist) (Schedl, 2008). My own definition, during many years was that it dealt with “understanding music understanding”<sup>1</sup> (Herrera et al., 2009).

I keep in my memory the excitement and feelings of “doing the right thing” when I made my mind up to write a paper for the then-called “First Symposium on Music Information Retrieval” held in Plymouth, Massachusetts in October 2000. I was then working in automatic audio music analysis (a topic that, surprisingly to today’s criteria, was not so popular then), in collaboration with Geoffroy Peeters in the IRCAM, and the context of music information retrieval perfectly matched our own goals. My intuition was, therefore, that the conference should not be missed, as it could change my life. I guess this was somehow what happened then. The paper was accepted (Herrera, Amatriain, Batlle, & Serra, 2000) and, after some time, it was reworked, expanded and updated as a journal paper (Herrera, Peeters, & Dubnov, 2003) and it became my most-cited journal article until very recently. But that was just one early landmark in this history.

The work presented in this thesis is then a personal account of my research in this field during a long lapse of time. To organize the content of this thesis, different options were considered, worked and discarded; I have finally kept the “temporal” metaphor of the ages in the development of a community of common research interests and practices (a research field, in other words). It should be noted that my perspective here is not that of an historian: this is not a history of MIR nor even of my research; it is just a personal interpretation or organization of problems, papers and research reports in which I felt deeply involved, in a way that might make sense (hopefully to other people different than me). Contrasting to most of the theses in the field, I will not talk about cutting-edge, state-of-the-art systems, I will not try to sell any system or algorithm. Some of those ideas, algorithms and systems reported here were, at its time, in such category and this feature will, indeed, be mentioned.

The “three ages of MIR” concept refers to the hypothesis that the evolution of the discipline, up until now, might be characterized as evolving through (or maybe revolving around?) three different stages: *the age of extractors*, *the age of semantic descriptors* and *the age of context-awareness*. This way of viewing the evolution of the field is a direct derivation of what was proposed long ago for digital media (Nack, 2004) but in this elaboration I am also getting some inspiration (surprisingly) from a taxonomy of intellectual behaviour levels used by educational psychologists<sup>2</sup> (Bloom et al., 1956;

---

<sup>1</sup> This concept is taken from Fiske (2008), who uses it in a different way and context.

<sup>2</sup> A figure with such taxonomy can be found here

<https://meestervormgever.wordpress.com/2015/02/05/revised-blooms-taxonomy-center-for-excellence-in->

Anderson & Krathwohl, 2001), as if the evolution of the discipline passed through similar stages as an educational process. The age of extractors was driven by the need to list, recognize and identify audio and music descriptors, and to develop audio and music algorithms to convert the signal into, mostly, numerical features. The drive of the age of semantics was the need to address summarization, clarification and prediction, leading the research efforts towards connecting the signal with users' mental representations of its content. Symbolic representations (categories and tags, for example, became part of the research topics then). Bringing "context" into the picture opened the possibility of integrating, differentiating or judging music and hence, in the age of context-awareness, researchers' attention turned on many aspects (environmental variables, cultural data sources, personal biographies, etc.) that were previously overlooked, ignored, or that were not addressable before, because of technological limitations or because of conceptual biases. The many different "contexts" could be found surrounding the user or listener, surrounding the music object that was on the focus of his/her interest, or framing his/her cultural beliefs and practices. A possible fourth age, not anticipated by Nack (2004), will be hypothesised as we reach the final chapters of the thesis. This one would be driven by operations such as generation and creation, meaning that users, content and metadata interact and influence one to another one, thanks to systems that go beyond the multi-level and multi-dimensional user-centred and contextualized analyses of music, which were achieved in the previous stages. These new systems will apply all that wealth of knowledge in music creation scenarios like digital audio workstations, DJ decks, sound installations or live concerts.

The observations, speculations and summaries on the MIR evolution that are contained here can be viewed as a continuation of other publications that I decided to be left out for conciseness: "The Discipline formerly known as MIR", presented in a special session on "The Future of MIR" carried out in ISMIR 2009 (Herrera et al., 2009), and "MIRrors: Music Information Research reflects on its future" (Herrera & Gouyon, 2013), written as the editorial of a special issue on that was published by the Journal of Intelligent Information Systems on the occasion of a special session in ISMIR 2012. Contrasting to such partial and subjective views, comprehensive technical introductions to MIR can be found in some good books on the topic (Lerch, 2012; Müller, 2015; Knees & Schedl, 2016; Weihs et al., 2016).

The core of all the forthcoming chapters in this thesis is formed by a selection of papers that I co-authored, and in which I played different roles (supervisor, idea generator, co-writer, statistics advisor, software tester/designer, or principal writer). Because my most important role was that of research supervisor, any error or mistake should be attributed to me in the first place instead of blaming any other person. There follow three chapters, each one centred on one of the above-mentioned "ages of MIR", then a fourth one where we discuss on the possibility that a new age could be happening and how this one would be characterized, and a final one containing closing thoughts. Each chapter (leaving aside the last one) starts with a short introduction to its topic, continues with the list of selected papers, and closes with a summary of the contributions made in the reported research. To provide a visual tool for getting some "big picture", figures 1.1, 1.2. and 1.3 show the

---

[learning-and-teaching/](#). It has not been included in the text as it only provides a loose connection with the main concepts of the chapter.



selected papers on a timeline where other landmarks (mostly projects framing the reported research) have been included.

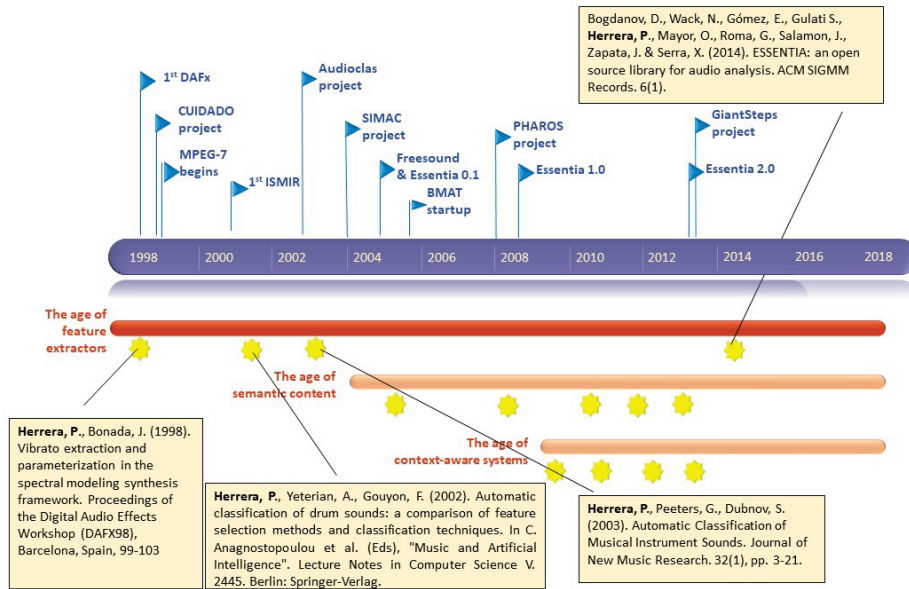


Figure 1.1 Timeline with the articles included in the age of feature extractors

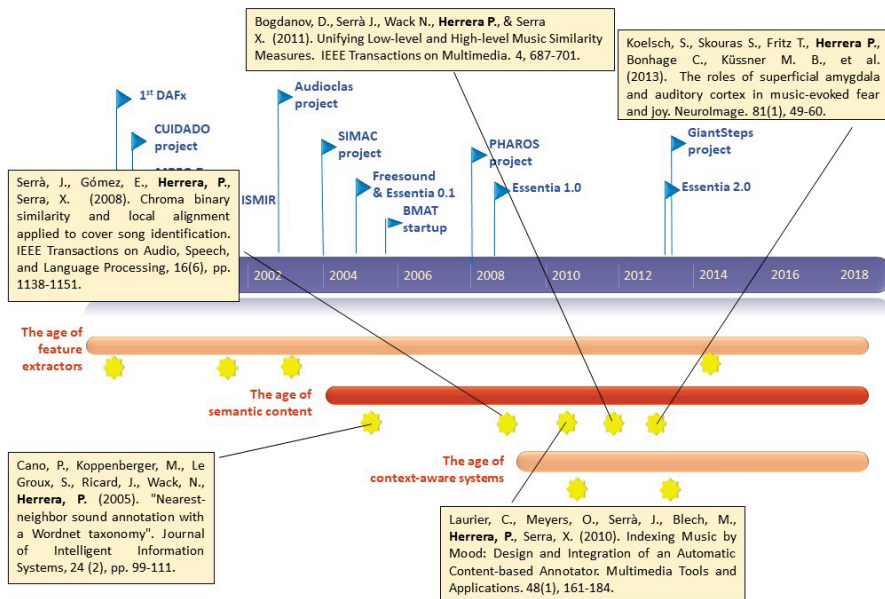


Figure 1.2 Timeline with articles in the age of semantic descriptors.

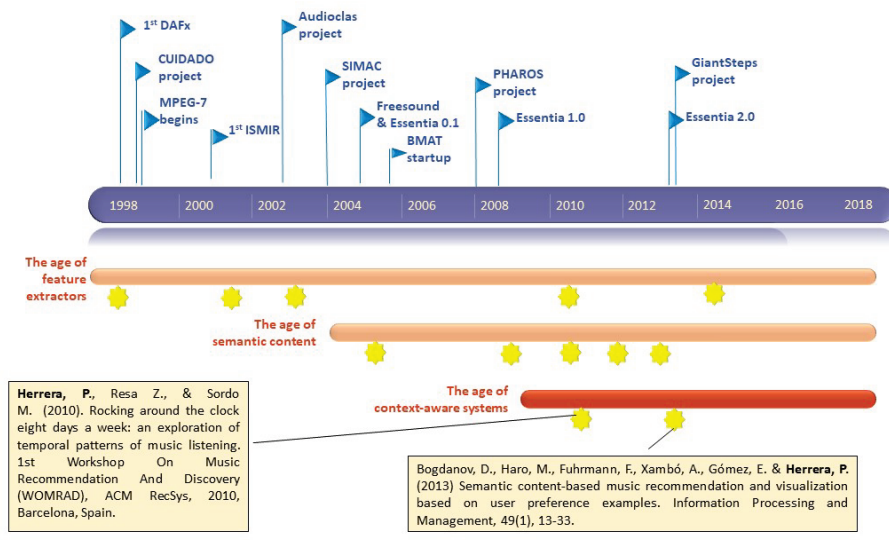


Figure 1.3. Timeline with articles in the age of context-aware systems

## 2. THE AGE OF FEATURE EXTRACTORS

*“That crystalline sound”, Jim [Morrison] jumped in. “I like the sound of broken glass falling from the void into creation.” “Which sound was that?” said Paul Beaver. “A couple back from where you are now,” Rothchild said. “It reminded me of the Kabbalah,” said Jim. “Kether the I AM, creating duality out of the one. All crystalline... and pure. You know, that sound.” “Did I make a sound like that?” “Sure,” Jim said. “A couple back.” And Paul Beaver began to unplug and replugin patch cords, and twist little knobs, and strike the keyboard, which emitted strange and arcane and utterly unearthly tones that sounded nothing like the Kabbalah or Kether; the crown of the Sefiroth. None of the sounds he was creating sounded pure and crystalline. And then we realized...he couldn't get back.*

Ray Manzarek, *Light My Fire: My Life with The Doors*.  
New York: Putnam, 1998, pp. 257.

*It's more fun to compute (x2)*

Ralf Hütter / Florian Schneider-Esleben / Karl Bartos

### 2.1. Introduction

It should not be surprising that the building blocks of MIR are features describing signal properties, close to what Nack (2004) termed “the physical surface” of the audio file or stream. Because of that proximity to the physical encoding these descriptors have been frequently considered “low-level features”. MIR would be unfeasible without the existence of music descriptors or features that grant decisions on differences or similarities between two or more music entities (audio excerpts, scores, etc.), or that allow such entities to be sorted or ranked. Signal processing techniques make feature extraction feasible and provide a solid ground for defining and developing a wealth of different algorithms that are capable of multi-featured descriptions (of properties related to amplitude, to frequency, to the spectral shape, to temporal changes, etc.) (Klapuri & Davy, 2006). Examples of such features can be the representation of a spectral envelope as a frequency bin vector output from an DFT, the representation of the energy by means of the Root Mean Square value, noisiness, fundamental frequency, or the almost ubiquitous Mel-frequency cepstral coefficients. In this chapter we provide an account of our ideas and some contributions that were aligned or shaped what we (following Nack, 2004) have called “the age of feature extractors”. Good summaries of features used for music content analysis can be found in recent textbooks (Lerch, 2012; Müller, 2105), in addition to the “classic” early account on them (Peeters, 2004).

As the creation of music has usually social functions and becomes the result of coordinated combination of different performers, a natural approach to the description of their recordings has involved source separation (Plumbley et al., 2002). Considering that human brains do not really “separate” sources into representations alike to those of a

multitrack recorder<sup>3</sup> (Bregman, 1990; Woods & McDermott, 2015; Puvvada & Simon, 2017; Hausfeld et al., 2018), and taking into account the complexities and errors of most of the source separation approaches, at least in the beginning of this age, we opted, from the very beginning for an approach termed *understanding without separation* (Martin et al., 1998)<sup>4</sup>. Under our view, features should be extracted from the raw audio without any previous attempt to separate instruments, melodic lines or noises and harmonics. In our long-term research program there was an explicit assumption that there is a lot of information available in the waveform and its low-level representations, and that this should make possible to compute relevant descriptors of what is going on there. Finding or inventing usable (and, indeed, computable) features, even if they were just rough approximations to musicians' or listeners' concepts, was the main goal during this age of feature extractors.

Most of the early research reported in this chapter was also motivated by our involvement in the development of a doomed standardization initiative that was named MPEG-7<sup>5</sup> and that attracted many people from companies and universities mostly between 1999 and 2001. Aligned with the age of feature extractors, MPEG-7 aimed to devise a way to automatically (or semi-automatically) describe any kind of multimedia content (more formally, a "Multimedia Content Description Interface")<sup>6</sup>. Audio and music description schemes were devised, discussed and adopted in meetings that were closer to business pitching than to scientific debating, and that were mostly driven by big companies' interests, instead of true and realistic considerations and requirements from potential users of those systems to be built upon such features and description schemes.

A third crucial aspect in the age of feature extractors is the shaping of a research field around the so-called "ISMIR"<sup>7</sup> conferences. I was lucky to get a paper accepted in the first of such conferences (Herrera et al., 2000), something that, at that time, I suspected it could change my research trajectory (as I think it did). That paper evolved into a journal article and was expanded to become one of those included in this chapter. ISMIR conferences have made possible, for the good and for the bad, to define methodological practices and research agendas, which is what characterizes a research community. In the beginning of the age of feature extractors special attention was devoted to defining figures of merit to objectively evaluate algorithms, or to creating the first annotated collections that facilitated such systematic evaluation. A practical consequence of that was the institution of the MIREX (Music Information Retrieval Evaluation eXchange)<sup>8</sup>, a yearly tournament where algorithms using cutting-edge features and processing architectures to solve specific problems such as chord detection, structural segmentation, tempo

---

<sup>3</sup> As literatura suggests, the typical representation is alike to figure plus background, though in cases of specific auditory training a bit more sophisticated multi-streaming can be observed.

<sup>4</sup> We just need to see what can be achieved with convolutional networks and other deep learning techniques (Uhlich et al., 2017; Ward et al., 2018) and with software such as Izotope RX or Celemony's Melodyne to realize how the topic and the technologies have progressed/improved in recent years.

<sup>5</sup> <http://mpeg7.org/mpeg-7-standard/>

<sup>6</sup> The number seven indicated a leap from the previous successful MPEG low bitrate coding of audio and video standards, known as MPEG-1 (where the, in another time ubiquitous, MP3 format was defined), MPEG-2, and MPEG-4.

<sup>7</sup> Originally "International Symposium on Music Information Retrieval", it soon became the "International Society for Music Information Retrieval" conference, reflecting the consolidation of a community of practitioners, with specific problem agendas and techniques on its own.

<sup>8</sup> [http://www.music-ir.org/mirex/wiki/MIREX\\_HOME](http://www.music-ir.org/mirex/wiki/MIREX_HOME)

estimation or drum transcription are compared and evaluated. The International Society of Music Information Retrieval<sup>9</sup> has been, since 2008, the formal representation of such growing community of multidisciplinary researchers with a common interest in music information retrieval.

A fourth factor defining the age of extractors were, probably, some of the music and computing research projects that the (then) European Commission funded. CUIDAD<sup>10</sup> and CUIDADO<sup>11</sup> (Content-based Unified Interfaces and Descriptors for Audio/music Databases available Online) were projects led by IRCAM between 1999 and 2003. These projects became the framework in which we developed our first set of descriptors to provide functionalities for an application called “The Sound Palette” (Vinet et al., 2002), which was intended to be a tool for metadata generation/annotation, in parallel with the generation of content. Some of those descriptors and ideas on how to organize flexible but powerful data structures to allow for music and audio queries became part of the above-mentioned MPEG-7 standard (Peeters, 2004; Peeters, et al., 2000; Herrera et al., 1999a; Herrera et al, 1999b). For historical rigour, it must be acknowledged that they were not the only European projects pushing in the direction of feature extraction at that time. Another remarkable influence was the work on machine listening that, in the decade surrounding the year 2000, was done in the MIT under the supervision of Barry Vercoe (Ellis, 1996; Casey, 1998; Martin, 1999; Scheirer, 2000; Smaragdis, 2001; Kim, 2003; Chai, 2005; Whitman, 2005; Jehan, 2005).

As we will discuss in the next chapter, low-level features are, nevertheless, not enough to allow the operations and functionalities required by the diverse research and application scenarios that MIR has progressively tackled (e.g., structural segmentation, classification, recommendation, transcription or creative transformation). Some of these applications have demanded new features and, consequently, the age of extractors has been overlapping with other ages in recent times. We are in fact still living in an age of extractors: even though they have been the initial quest of MIR researchers, and even though their sophistication, coverage and effectiveness has increased along the two decades here contemplated, they are still a fertile ground for research (Peeters et al., 2015). Proposals such as I-Vectors (Eghbal-zadeh, Lehner et al., 2015, Eghbal-zadeh, Schedl et al., 2015), which adapt to music a state-of-the-art technique for speaker verification, are getting increasing attention (Park et al., 2018). Features inspired in auditory perception seem to be still a fruitful path too (Richard et al., 2013; Hemery & Aucouturier, 2015). Finally, features extracted from hidden layers of deep networks, as naturally as with convolutional neural networks (LeCun et al., 1998; Humphrey, Bello and LeCun, 2013; Pons et al., 2017), or using other architectures and data interpretation techniques (Lee et al., 2009; Hamel & Eck, 2010; Schmidt & Kim, 2013; Dieleman & Schrauwen, 2014; Kereliuk & Sturm, 2015), are the most promising path to break the glass-ceiling reported in most of the typical music description problems (Aucouturier & Pachet, 2004; Pampalk, Flexer and Widmer, 2005). These abstract features show also properties that grant transfer learning (i.e., to be directly applied to different conceptual problems other than the one in which they have been created) (Choi et al., 2017b).

---

<sup>9</sup> <https://www.ismir.net/>

<sup>10</sup> [https://cordis.europa.eu/result/rcn/26464\\_en.html](https://cordis.europa.eu/result/rcn/26464_en.html)

<sup>11</sup> [https://cordis.europa.eu/project/rcn/57197\\_en.html](https://cordis.europa.eu/project/rcn/57197_en.html)

## 2.2. Papers included in this chapter<sup>12</sup>

**Herrera, P.** & Bonada, J. (1998). *Vibrato extraction and parameterization in the spectral modelling synthesis framework*. Proceedings of the Digital Audio Effects Workshop (DAFX98), Barcelon, Spain. (paper cited 74 times)

**Herrera, P.**, Yeterian, A., Gouyon, F. (2002). Automatic classification of drum sounds: a comparison of feature selection methods and classification techniques. In C. Anagnostopoulou et al. (Eds), "Music and Artificial Intelligence". Lecture Notes in Computer Science V. 2445. Berlin: Springer-Verlag. (Series IF: 0.8; Q2 in Computer Science journals; paper cited 148 times)

**Herrera, P.**, Peeters, G., Dubnov, S. (2003). Automatic Classification of Musical Instrument Sounds. *Journal of New Music Research*. 32(1), pp. 3-21. (Journal h-index: 22; Journal IF 2016: 1.122; Q1 in music-related journals; paper cited 231 times)

Gómez, E. & **Herrera, P.** (2008). Comparative Analysis of Music Recordings from Western and Non-Western traditions by Automatic Tonal Feature Extraction. *Empirical Musicology Review*, 3(3), pp. 140-156. (paper cited 33 times)

Bogdanov, D., Wack, N., Gómez, E., Gulati S., **Herrera, P.**, Mayor, O., Roma, G., Salamon, J., Zapata, J. & Serra, X. (2014). ESSENTIA: an open source library for audio analysis. *ACM SIGMM Records*. 6(1). (Winner of ACM MM 2013 Open Source competition; 5 citations, but a longer report of Essentia (Bogdanov et al., 2013a), not from a journal, has been cited 204 times)

## 2.3. Contributions in the selected papers

Since 1998, I have been contributing to develop, test and disseminate acoustic features that are specifically tuned to music description problems. Even though it is not a journal paper, I have included here, because it is still frequently cited, our early and therefore somehow pioneer paper “Vibrato extraction and parameterization in the spectral modelling synthesis framework” (presented in the very first DAFX conference, in which organization I also collaborated). In that paper we proposed a couple of strategies for detecting expressive frequency modulations (i.e., vibrato) in monophonic audio. As a property of fundamental frequency, vibrato could be detected, characterized (rate and depth) and even suppressed (up to a certain point) by means of either low-pass filtering the F0 envelope or by applying an IFFT to it, once the time-varying F0 was extracted using a “harmonic plus noise” (Serra, 1989; Serra et al., 1997) decomposition. Surprisingly to our current standards, but being the norm in those early days, the proposed approaches were justified by discussing just a few of positive examples (and sometimes using some failures or errors as a source for discussion or indicating further required work). Annotated collections and large-scale evaluations, including statistics and figures of merit, were still outside the common practices in the dawn of the age of feature extractors.

---

<sup>12</sup> Citation count retrieved from Google Scholar, on September 10<sup>th</sup>, 2018, for all the selected papers in this thesis.

In “Automatic classification of drum sounds: a comparison of feature selection methods and classification techniques” (Herrera et al. 2002) we studied classifiers and features that discriminated well the isolated sounds of a drum kit. This was, to my knowledge, one of the earliest papers using Weka<sup>13</sup> for building music-related classification models. At that time, addressing music description problems with machine learning was not that usual as it has been since then. As far as I know, the paper contained the first proposal of a generic method for the automatic classification of isolated drum sounds (i.e., telling apart snares from toms, from bass drums, from hi-hats, etc.) using features computed from the audio, and also explored the differences between flat classification and hierarchical classification involving first the determination of the subfamily of the instrument (i.e., membranes versus plates). We kept on working on this topic for some time, using different approaches and application scenarios (Herrera et al., 2003; Sandvold et al., 2004; Pampalk et al., 2004; Herrera et al., 2004; Pampalk et al., 2008; Haro & Herrera, 2009; Gómez-Marín et al., 2017).

In "Automatic Classification of Musical Instrument Sounds" (Herrera et al., 2003) we further elaborated what had been our contribution to the first ISMIR conference (Herrera et al., 2000). This has been my most cited work until recently, probably because of its tutorial tone, as it basically summarizes the state of the art (at that time) on the topic of sound descriptors and sound classification (as a matter of curiosity, it was one of the earliest papers remarking the potential of Support Vector Machines for classification, when very few papers on music audio analysis were using them)<sup>14</sup>. Although other feature taxonomies were proposed in that period and have probably become more successful (Lessafre et al., 2003; Peeters, 2004a; Lessafre, 2006), we proposed a classification of features according to four points of view:

1. The *temporal dynamics* of the computed feature, i.e. the fact that the features represent a value extracted from the signal at a given time (instant, static), or a parameter from a model of the signal behaviour along time (mean, standard deviation, derivative or Markov model of a parameter);
2. The *temporal scope* of the description provided by the features: some descriptors apply to only part of the object (e.g. description of the attack of the sound), whereas other ones apply to the whole signal (e.g. loudness). To this respect, (Snyder, 2000) discusses three different time levels of musical experience in human listeners: event fusion happens at the shortest level (less than 30 milliseconds); melodic and rhythmic grouping happens between 0.03 and 8 seconds; finally, *form* elements require more than 8 seconds. Considering neurophysiologic evidence, it seems that feature extraction (pitch height and chroma, intensity, roughness, timbre...) takes not more than 200ms, melodic, rhythmic and tonal analysis takes not more than 400ms, meaning is available after 500ms of processing, and basic structural building and reanalysis takes up to 900ms (Koelsch & Siebel, 2005). Even though these time

---

<sup>13</sup> Weka (Witten et al., 2016) was one of the most popular machine learning toolboxes during the first decade of the 21st Century, and it is still frequently used as proved by the recent current fourth edition of the textbook associated to it.

<sup>14</sup> A somehow “companion” to that paper, also with a tutorial-like organization, was the book chapter “Automatic classification of pitched musical instrument sounds” (Herrera, et al., 2006).

constraints are not closely followed by our artificial systems, the order and magnitude of the corresponding computational complexity can be considered as equivalent.

3. The “*abstractness*”, i.e. what the feature represents (e.g. cepstrum and linear prediction are two different representation and extraction techniques for representing spectral envelope, but probably the former one can be considered as more abstract than the latter, as the numbers that are computed do not have a direct interpretation);
4. The *extraction process* of the feature. According to this point of view, we could further distinguish:
  - Features directly computed on the waveform data as, for example, zero-crossing rate (the rate that the waveform changes from positive to negative values), temporal centroid, log-attack time, or amplitude RMS;
  - Features extracted after performing a transform of the signal (FFT, wavelet...) as, for example, spectral centroid (i.e., the “gravity centre” of the spectrum), spectral flatness, or the harmonic pitch class profile (HPCP);
  - Features that relate to a signal model, as for example the sinusoidal model or the source/filter model or the coefficients of a Mel-cepstral transform;
  - Features that try to mimic the output of the ear system (Bark or ERB filter outputs).
  - Features that require the application of a class model (i.e., genre, instrumentation, etc.), usually becoming “semantic” features (see next chapter).

In "Comparative Analysis of Music Recordings from Western and Non-Western traditions by Automatic Tonal Feature Extraction" we raised an early warning of the need to address the study of music cultures that, in the history of computer music in general, and in the early years of MIR were totally overlooked. As many musical cultures in the world do not use any kind of notation to preserve and transmit their heritage, audio features seemed the right vehicle to study their similarities and differences, if only with a very broad, generic, naïve and even simplistic perspective. In this article we wondered if we could use some of our available descriptors to compare music from different cultures. We analysed which descriptors were the most relevant to distinguish between music from broadly defined cultures (i.e. Africa, Java, Japan, India...), represented by a collection of over 1500 audio files, and observed that such discrimination could be on the basis of the shapes in the data distributions of certain features (such as tuning, deviations from equal temperament, statistics of low-level tonal descriptors, dissonance or timbre-related descriptors), achieving accuracies above 85%. Some years later, the most ambitious MIR project to bring important musical traditions to the focus of MIR researchers, COMPMUSIC, was launched in our group<sup>15</sup> (Serra, 2011). I dare to speculate here that the work done for our paper was among the inspirational sources that led to the COMPMUSIC project proposal.

Wrapping up more than a decade of working on feature extraction, “ESSENTIA: an open source library for audio analysis” (Bogdanov et al., 2014) reports on the library that has been developed and used in many research projects carried out in the MTG and all over the world (see also Bogdanov et al., 2013a and Bogdanov et al., 2013b). The library runs in the three main computing platforms and makes possible the computation of timbre, loudness, pitch, rhythm, tonal and morphological descriptors, in addition to their

---

<sup>15</sup> <http://compmusic.upf.edu/>



statistical moments. It also includes Python bindings and Vamp<sup>16</sup> plugins for easy extension, integration and prototyping. The impact that Essentia created (and is still creating) on the research community, can be somehow tracked in a page<sup>17</sup> that lists the papers using it. Many of them have not been done inside or in connection with the MTG and, because of such wide adoption, the library is still maintained and periodically updated.

As mentioned before, the age of feature extraction coexists with other ages, because no initial problem can be considered “solved” yet, and newer features have not exhaustively covered the huge space of perspectives that can be adopted when describing waveforms and its content. A new turn of the road has been witnessed, for example, since the introduction of deep learning algorithms. Now, researchers seem to be confident on the computing capabilities of hidden-layered networks to “discover” the right features to effectively model a problem (Choi et al., 2017a; Choi et al., 2017b; Han et al., 2017; Lee & Nam, 2017; Pons et al., 2017). Although the promise of automatic feature discovery can be traced back to a system such as Pachet’s Extractor Discovery System (Zils & Pachet, 2004), there is an important tradeoff involved in it: interpretability. There are indeed many scenarios where interpretability is not important, provided that the system performs top-notch. But domain knowledge brought in by humans, even in such situations, can be a bonus to properly crafting such systems, not to mention the Occam’s razor or parsimony principle that has guided human research as we have known in the past centuries (i.e., an algorithm that is orders of magnitude more complex than another one should provide a performance that is orders of magnitude better than the simpler one).

---

<sup>16</sup> <https://www.vamp-plugins.org/>

<sup>17</sup> [http://essentia.upf.edu/documentation/research\\_papers.html](http://essentia.upf.edu/documentation/research_papers.html)

**Herrera, P. & Bonada, J. (1998).** “Vibrato extraction and parameterization in the spectral modelling synthesis framework”. Proceedings of the Digital Audio Effects Workshop (DAFX98).

<http://mtg.upf.edu/system/files/publications/Herrera-DAFX-1998.pdf>

# Vibrato Extraction and Parameterization in the Spectral Modeling Synthesis Framework

Perfecto Herrera, Jordi Bonada  
Audiovisual Institute, Pompeu Fabra University  
Rambla 31, 08002 Barcelona, Spain  
{pherrera, jboni}@iaa.upf.es <http://www.iaa.upf.es>

## Abstract

Periodic or quasi-periodic low-frequency components (i.e. vibrato and tremolo) are present in steady-state portions of sustained instrumental sounds. If we are interested both in studying its expressive meaning, or in building a hierarchical multi-level representation of sound in order to manipulate it and transform it with musical purposes those components should be isolated and separated from the amplitude and frequency envelopes. Within the SMS analysis framework it is now feasible to extract high level time-evolving attributes starting from basic analysis data. In the case of frequency envelopes we can apply STFTs to them, then check if there is a prominent peak in the vibrato/tremolo range and, if it is true, we can smooth it away in the frequency domain; finally, we can apply an IFFT to each frame in order to re-construct an envelope that has been cleaned of those quasi-periodic low-frequency components. Two important problems nevertheless have to be tackled, and ways of overcoming them will be discussed in this paper: first, the periodicity of vibrato and tremolo, that is quite exact only when the performers are professional musicians; second: the interactions between formants and fundamental frequency trajectories, that blur the real tremolo component and difficult its analysis.

## 1 Introduction

Long sustained notes become boring and uninteresting if their steady states have a strictly constant fundamental frequency. Because of that and other musical reasons to be found in music performance treatises, good performers invest a lot of time to developing proficiency in techniques for the continuous modulation of frequency and/or amplitude. This kind of modulations are respectively called *vibrato* and *tremolo* and its feasibility for every instrument depends on its sound generation mechanisms (for example, string instruments favor vibrato and otherwise reeds favor tremolo). The scientific study of vibrato can be traced backwards to the work by Seashore [1] who, notwithstanding his technological limitations, yielded a rough but valid characterization of that phenomenon. More recent studies ([2], [3], [4], [5]) have been developed with the help of modern analysis techniques and devices but we can conclude that, although we understand the basic facts about vibrato in different musical instruments, more research is needed on vibrato as a physical phenomenon (not to mention as a musical resource, indeed), specially on its temporal evolution and its way of change between consecutive notes. Anyway, if it is parametrically and/or procedurally possible to describe vibrato, it should be possible to manipulate it for musical, engineering, or acoustical purposes.

As Desain and Hoenig [6] noted, the shape of musically modulated signals is quite complex to be extracted without a solid model of analysis. One of those models could be the Spectral Modeling Synthesis (SMS) developed by Serra [7]. Recent software developments inside that framework ([8], [9], [10]) have made possible to segment a continuous signal such as a musical phrase or a long note into different *regions* that have different basic *features* or *parameters* both static and evolving along time (i.e. mean fundamental frequency, mean amplitude, amplitude tendency, noise profile, amplitude and fundamental frequency envelope, etc.); once those parameters have been extracted, it is possible to manipulate them separately in order to achieve delicate sound transformations during the re-synthesis stage. Consequently there are certain situations in which it could be useful to separate the contribution of modulation processes over a stable set of parameters in order to achieve a greater flexibility and better quality of synthesis and transformation.

The vibrato problem can be decomposed into three subproblems to be tackled: 1) Identification (or Detection and Parameterization); 2) Extraction; and 3) Re-synthesis. Considering that our target system is an off-line (non real-time) one in this paper we will focus on the first two points (see [11] for a synthesis oriented paper).

## 2 Frequency –domain strategy

From the frequency-domain point of view, vibrato detection in an off-line system assumes that a steady state has been correctly delimited and parameterized in a previous stage of the analysis; that is to say that we have obtained a fundamental frequency track whose frequency is constrained in a range of less than a whole tone around an ideal “mean” (although the usual vibrato depth reported by different studies carried away with professional musicians is lesser than half a tone, it should be noted that not so well trained performers generate larger excursions from the nominal fundamental frequency). The fundamental frequency track obtained in SMS analysis is an envelope of data representing Hertz along time, and has a number of values equal to the frame rate of analysis (typically we use 345 points per second so that each one of our *envelope frames* integrates information for such a temporal lapse); thus, that envelope will be the starting data for the process (for details see [9]).

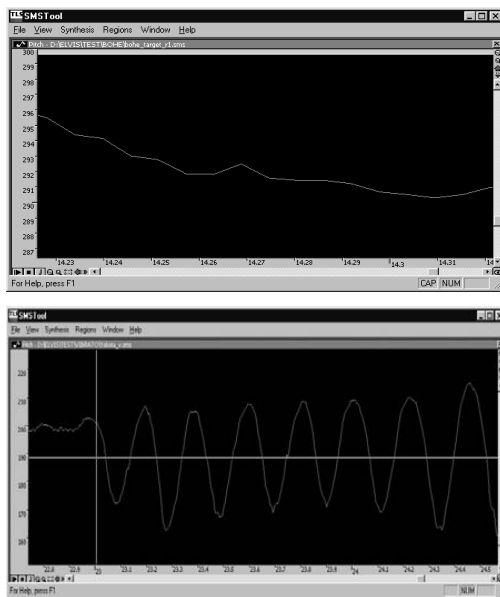


Figure 1. Two fundamental frequency tracks: a) from a steady state portion of sound without vibrato; b) from a steady state portion of a sound with vibrato.

The vibrato detection proceeds as follows: the discrete fundamental frequency track is first transformed into a 0-centered track by computing the global mean and subtracting it from every fundamental frequency value in the original track. This smoothed and 0-centered track is then windowed. A window size of 128 points or 0.37 seconds (more than 2 times the lowest period that is

expected to be found for a vibrato) with a 50% of window overlap has been proved suitable for our purposes. Different kind of windows (Blackman-Harris, Kaiser, Hamming) have been tested yielding no significant differences. For each window its FFT is then computed and spectral peaks are calculated with parabolic interpolation. In case the analyzed region has vibrato we get a prominent peak around 5/6 Hz (in fact it is the most prominent peak detected). As expected, such a clear and stable peak is not present when the region has no vibrato. The vibrato detection process concludes with the extraction and storage of the rate and the depth of vibrato as high-level parameters of the analyzed frame (in fact, they will be later pooled with the values for all the *envelope frames* and global mean values will be extracted for a whole region).

At this point, the vibrato extraction proceeds. Different algorithms could be implemented, as for example a similar one to the SMS low-level analysis (i.e. by additive synthesis and subtraction of the harmonic part), but it is more economic and easy to “crop” the prominent peak (and sometimes the second one) of every envelope. Then the IFFT of the altered spectrum is computed so that we get a signal without the modulation components, that is, more stationary than the original one.

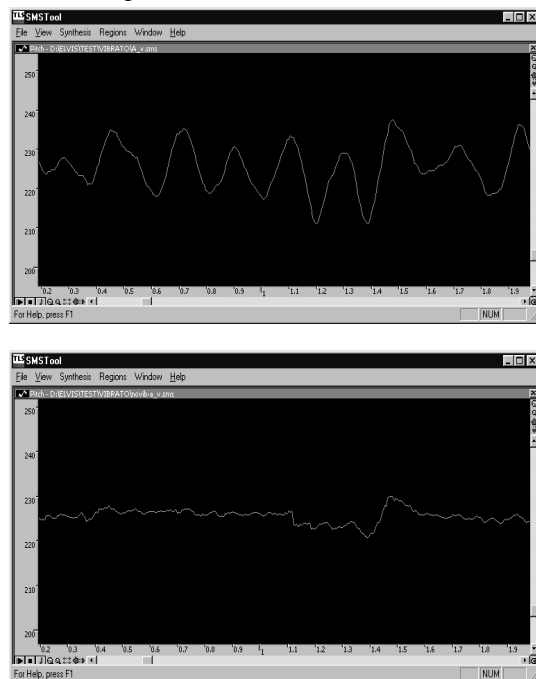


Figure 2. Comparison of a fundamental frequency track of a steady state portion of sound: a) with its original vibrato; b) the same fundamental track after vibrato extraction in the frequency-domain.

### 3 Time-domain strategy

In the time-domain there are several robust techniques for fundamental frequency estimation [12] that could be suitable for vibrato extraction. Besides that, time-domain strategies offer important advantages such as the option of using shorter windows. In such a scenario, we could find practical situations that only demand to get rid of vibrato, but not necessarily to characterize it in full detail. Given that constraint, a filtering strategy seems quite suitable to be approached (on the other hand, see [13] for a time-domain -although not real-time- complete solution without using filters).

In order to design an appropriate filtering algorithm for this task we have to take into account the fact that the value to be given to the filter at every point of time should be centered around a conventional 0 (in this case the mean fundamental frequency). For an off-line system such a central value could be the mean frequency of the steady state, but in a real-time system that center must be approximated from the past behaviour of the fundamental frequency track. If much previous information is used in computing such an approximation then we will lose the temporal trends for the pitch, but if only very recent values are taken into account then we will lose the low frequency modulations that we are addressing to (in fact this is something like a paradox, because losing low frequency modulations is what we are trying to achieve!). An acceptable solution should be a number of points spanning more than a common vibrato cycle, and at least one of the shortest vibrato cycle we could find. After some trial and error we settled into a filter buffer that takes into account the preceding 80 envelope points (about one vibrato cycle of 4.5 Hz) and does not blur the mid-term variations of the fundamental. If the system does not yet have 80 data points it uses the mean of the available points. We feel, nevertheless, that this mechanism should be exhaustively refined in order to obtain better results as we can see from *Figure 3*.

After we get the discussed mean value, we can apply a filter to the incoming data. Because both the vibrato rate and depth will be constrained, we have implemented a 6-order Butterworth high-pass filter that effectively eliminates frequencies lower than 10 Hz from the fundamental frequency track. The selection of the filter was done with the help of MATLAB, and finally we opted for a filter defined along the following parameters: (passband=.25 radians (approx. 21Hz.), stopband=.11 radians (approx. 9 Hz), passband ripple = 3 dB, stopband attenuation = 40 dB).

Although this strategy does not allow to characterize vibrato at the filtering stage, the blackboard-like

model implemented in the SMS analysis framework facilitates that vibrato parameters can be extracted later on by picking the relevant information from other concurrent analysis modules (of course there is an arguable time-resolution payoff).

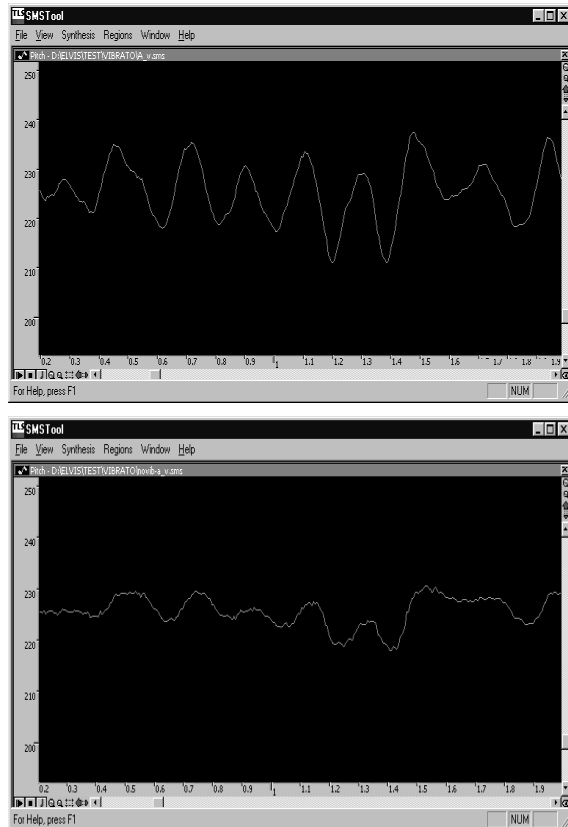


Figure 3. Comparison of an original fundamental track of a steady state portion of sound: a) with original vibrato; b) after time-domain vibrato extraction.

### 4 Interaction between vibrato and spectral shape

If we examine the amplitude track of a region with vibrato it seems that there also are cyclic modulations around an ideal “mean value”. Although it could be tempting to consider them as examples of a concomitant “tremolo” and therefore to proceed with that track as we did with the frequency, we should be warned that superficially similar expressive resources as vibrato and tremolo could have different musical meanings and uses, and do not need to be associated. It should also be noted that (at least in human singing) amplitude variations follow a pattern not as regular as frequency does. In fact the main factor for the observed variations in amplitude are, other than a

tremolo process, the interaction between the vibrato process and the resonances of the vocal tract [14], [15]. Therefore, our strategy for eliminating those amplitude modulations goes as follows: in the frequency-domain case, an spectral envelope for the steady region is computed; then, we proceed by recalculating the “right” amplitude value for every track (or partial) frame by frame. By “right amplitude” we mean the amplitude that the track should have, considering the trajectory correction induced by the vibrato-suppression procedure (for example, let’s suppose that the original fundamental frequency track entered a resonance region; after vibrato suppression its amplitude would be still reflecting its presence in such a resonance region, but in fact the track is not there anymore, so we will interpolate –from the spectral envelope– the amplitude corresponding to the current spectral location for that track).

On the other hand, in the time-domain case we are just starting to implement an “incremental-resolution spectral envelope extracting algorithm” much in the vein of the spectral tracings used by [11], and similar to the one apparently used by humans [16]. Such an algorithm, whereby the correction of the amplitudes is done frame by frame (as explained before), computes a spectral envelope that gets increasing resolution as we get more frames from the basic analysis.

## 5 Conclusions

In this paper we have presented a two-fold approach for managing vibrato inside an specific analysis/synthesis framework like SMS. Although the higher level attributes extracted in the analysis process allow both the satisfactory characterization of vibrato and also its removal from a steady state portion of a sound in the frequency domain, there will be practical situations in which only removal will be mandatory and then we can apply a simpler time-domain strategy. Nonetheless more research is needed, and it shall be pursued for us, in order to refine the current algorithms, and, afterwards, achieve a flexible and acceptable synthesis of vibrato notes.

Sound examples related to this paper can be found at: <http://www.iaa.upf.es/~perfe/papers/dafx98poster-soundexamples.html>.

## References

- [1] C. E. Seashore. *Psychology of Music*. New York: McGraw-Hill, 1938. (Reprint: Dover, New York, 1967).
- [2] J. Sundberg. “Vibrato and vowel identification”. *Arch. Acoust.* 2, 257-266, 1977.
- [3] E. Prame. “Measurements of the vibrato rate of ten singers”. *J. Acoust. Soc. Am.* 96 (4), pp. 1979-1984, 1994.
- [4] H. Honing. “The vibrato problem, comparing two solutions”. *Computer Music Journal*, 19 (3), 1995.
- [5] P. Desain and H. Honing. “Towards algorithmic descriptions of continuous modulations of musical parameters”. *Proceedings of the ICMC*, 1995.
- [6] P. Desain and H. Honing. “Modeling continuous modulations of music performance”. *Proceedings of the ICMC*, 1996.
- [7] X. Serra. *A System for Sound Analysis/Transformation/Synthesis based on a Deterministic plus Stochastic Decomposition*. Ph.D. Dissertation, Stanford University, 1989.
- [8] X. Serra and others. “Integrating Complementary Spectral Models in the Design of a Musical Synthesizer”. *Proceedings of the ICMC*, 1997.
- [9] X. Serra and J. Bonada. “Sound Transformations Based on the SMS High Level Attributes”. *Proceedings of the Digital Audio Effects Workshop (DAFX98)*, 1998.
- [10] P. Cano. “Fundamental Frequency Estimation in the SMS Analysis”. *Proceedings of the Digital Audio Effects Workshop (DAFX98)*, 1998.
- [11] R. Maher and J. Beauchamp. “An investigation of vocal vibrato for synthesis”. *Applied Acoustics*, 30, pp. 219-245, 1990.
- [12] W. Hess. *Pitch determination of speech signals*. Berlin: Springer-Verlag, 1983.
- [13] S. Rossignol and others. “Feature Extraction and Temporal Segmentation of Acoustic Signals”. *Proceedings of the ICMC*, 1998.
- [14] Y. Horii. “Acoustic analysis of vocal vibrato: a theoretical interpretation of data”. *J. of Voice*, 3 (1). 36-43. 1989.
- [15] M. Mellody and G. H. Wakefield. “Modal distribution analysis of vibrato in musical signals”. *Proceedings of SPIE Conf.*, 1998.
- [16] J. H. Ryalls and P. Lieberman. “Fundamental frequency and vowel perception”. *J. Acoust. Soc. Am.* 72 (5). 1631-1634, 1982.

**Herrera, P.**, Yeterian, A., Gouyon, F. (2002). Automatic classification of drum sounds: a comparison of feature selection methods and classification techniques. In C. Anagnostopoulou et al. (Eds), "Music and Artificial Intelligence". Lecture Notes in Computer Science V. 2445. Berlin: Springer-Verlag.

DOI [https://doi.org/10.1007/3-540-45722-4\\_8](https://doi.org/10.1007/3-540-45722-4_8)

ISSN 0302-9743

ISBN 978-3-540-44145-8

Online ISBN 978-3-540-45722-0





# Automatic classification of drum sounds: a comparison of feature selection methods and classification techniques

Perfecto Herrera, Alexandre Yeterian, Fabien Gouyon

Universitat Pompeu Fabra  
Pg. Circumval.lació 8  
08003 Barcelona, Spain  
+34 935422806  
{pherrera, ayeter, fgouyon}@iua.upf.es

**Abstract.** We present a comparative evaluation of automatic classification of a sound database containing more than six hundred drum sounds (kick, snare, hihat, toms and cymbals). A preliminary set of fifty descriptors has been refined with the help of different techniques and some final reduced sets including around twenty features have been selected as the most relevant. We have then tested different classification techniques (instance-based, statistical-based, and tree-based) using ten-fold cross-validation. Three levels of taxonomic classification have been tested: membranes versus plates (super-category level), kick vs. snare vs. hihat vs. toms vs. cymbals (basic level), and some basic classes (kick and snare) plus some sub-classes –i.e. ride, crash, open-hihat, closed hihat, high-tom, medium-tom, low-tom- (sub-category level). Very high hit-rates have been achieved (99%, 97%, and 90% respectively) with several of the tested techniques.

## 1. INTRODUCTION

Classification is one of the processes involved in audio content description. Audio signals can be classified according to miscellaneous criteria. A broad partition into speech, music, sound effects (or noises), and their binary and ternary combinations is used for video soundtrack descriptions. Sound category classification schemes for this type of materials have been recently developed [1], and facilities for describing sound effects have even been provided in the MPEG-7 standard [2]. Usually, music streams are broadly classified according to genre, player, mood, or instrumentation. In this paper, we do not deal with describing which instruments appear in a musical mixture. Our interest is much more modest as we focus only in deriving models for discrimination between different classes of isolated percussive sounds and, more specifically in this paper, of acoustic “standard” drum kit sounds (i.e. not electronic, not Latin, not brushed, etc.). Automatic labelling of instrument sounds has some obvious applications for enhancing sampling and synthesis devices’ operating systems in order to help sound designers to categorize (or suggesting names for) new patches and samples. Additionally, we assume that some outcomes of research on this subject will be used for the more ambitious task of describing the instrumentation in a musical recording, at least of the “rhythm loop” type.

Previous research in automatic classification of sound from music instruments has focused in instruments with definite pitch. Classification of string and wind instrument sounds has been attempted using different techniques and features yielding to varying degrees of success (see [3] for an exhaustive review). Classification of percussive instruments, on the other hand, has attracted little interest from researchers. In one of those above cited studies with pitched sounds, Kaminskyj [4] included three pitched percussive categories (glockenspiel, xylophone and marimba) and obtained good classification results (ranging from 75% to 100%) with a K-NN algorithm. Schloss [5] classified the stroke type of congas using relative energy from selected portions of the spectrum. He was able to differentiate between high-low sounds and open, muffled, slap and bass sounds. Using a K-means clustering algorithm, Bilmes [6] also was able to differentiate between sounds of three different congas. McDonald [7] used spectral centroid trajectories as classificatory features of sounds from percussive instruments. Sillanpää [8] used a representation of spectral shape for identification of the basic five categories of a drum kit: bass drum, snares, toms, hihats, and cymbals. His research was oriented towards transcription of rhythm tracks and therefore he additionally considered the case of identification of several simultaneous sounds. A database of 128 sounds was identified with 87% of accuracy for the case of isolated sounds. Performance dramatically dropped when there were two or three simultaneous sounds (respectively 49% and 8% for complete identification, though at least one of the sounds in the mixture was correctly identified all the times). In a subsequent study [9], the classification method used energy, Bark-frequency and log-time resolution spectrograms, and a fuzzy-c clustering of the original feature vectors into four clusters for each sound class. Weighted RMS-error fitting and an iterative spectral subtraction of models was used to match the test sounds against learnt models. Unfortunately, no systematic evaluation was presented this time. Goto and Murakoa [10] also studied drum sound classification in the context of source separation and beat tracking [11]. They implemented an “energy profile”-based snare-kick discriminator, though no effectiveness evaluation was provided. As a general criticism, in the previous research there is a lack of systematic evaluation of the different factors involved in automatic classification, and the databases are small to draw robust conclusions. A more recent study on this subject in the context of basic rhythmic pulse extraction [12] intended to be systematic, but also used a small database and a reduced set of descriptors.

Research on perceptual similarity of sounds is another area that provides useful information for addressing the problem of automatic classification of drum sounds. In perceptual studies, dis-similarity judgments between pairs of sounds are elicited from human subjects. With multidimensional scaling techniques, researchers find the dimensions that underlie to dis-similarity judgments. Even further, with proper comparison between those dimensions and physical features of sounds, it is possible to discover the links between perceptual and physical dimensions of sounds [13], [14], [15], [16]. A three dimensional perceptual space for percussive instruments (not including bells) has been hypothesized by Lakatos [17] (but also see [18]). This percussive perceptual space spans three related physical dimensions: log-attack time, spectral centroid and temporal centroid. Additional evidence supporting them has been gathered during the multimedia content description format standardization process (MPEG-7) and, consequently, they have been included in MPEG-7 as

descriptors for timbres [19]. Graphical interactive testing environments that are linked to specific synthesis techniques [20] seem to be a promising way for building higher-dimensional perceptual spaces.

From another area of studies, those focusing on characteristics of beaten objects, it seems that information about the way an object is hit is conveyed by the attack segment, whereas the decay or release segment conveys information about the shape and material of the beaten object [21]. Repp [22] found that different hand-clapping styles (palm-to-palm versus fingers-to-palm) correlated with different spectral envelope profiles. Freed [23] observed that the attack segment conveyed enough information for the subjects to evaluate the hardness of a mallet hit. Four features were identified as relevant for this information: energy, spectral slope, spectral centroid and the time-weighted average centroid of the spectrum. Kaltzky et al. [24] have got experimental results supporting the main importance of the decay part (specifically the decay rate) of a contact sound in order to identify the material of the beaten object.

In the next sections we will present the method and results of our study on automatic identification of drum sounds. First we will discuss the features we initially selected for the task and the ways for using the smallest set without compromising classification effectiveness. Some techniques consider relevance of descriptors without considering the classification algorithm in which they are being issued, but there are also attribute selection techniques that are linked to specific classification algorithms. We will compare both approaches with three different classification approaches: instance-based, statistical-based, and tree-based. Classification results for three taxonomic levels (super-category, basic level classes, and sub-categories) of drum-kit instruments will then be presented and discussed.

## **2. METHOD**

### **2.1 Selection of sounds**

A database containing 634 sounds was set up for doing this study. Distribution of sounds into categories is shown in Table 1. Sounds were drawn from different commercial sample CD's and CD-ROMs. The main selection criteria were that they belonged to acoustic drums with as little reverberation as possible, and without any other effect applied to them. Also different dynamics and different physical instruments were looked for. Specific playing techniques yielding dramatic timbral deviations from a "standard sound" such as brushed hits or rim-shots were discarded.

**Table 1. Categories used and number of sounds (inside parentheses) included in each category**

Super-category	Basic-level	Sub-category
Membranes (380)	Kick (115)	Kick (115)
	Snare (150)	Snare (150)
	Tom (115)	Low (42)
		Medium (44)
	High (29)	
Plates (263)	Hihat (142)	Open (70)
		Closed (72)
	Cymbal (121)	Ride (46)
		Crash (75)

## 2.2 Descriptors

We considered descriptors or features belonging to different categories: attack-related descriptors, decay-related descriptors, relative energies for selected bands and, finally, Mel-Frequency Cepstral Coefficients and variances. An amplitude-based segmentator was implemented in order to get an estimation of the attack-decay boundary position, for then computing those descriptors that used this distinction. Analysis window size for the computation of descriptors was estimated after computation of Zero-Crossing Rate.

### 2.2.1 Attack-related descriptors

Attack Energy (1), Temporal Centroid (2), which is the temporal centre of gravity of the amplitude envelope, Log Attack-Time (3), which is the logarithm of the length of the attack, Attack Zero-Crossing Rate (4), and TC/EA (5), which is the ratio of the Temporal Centroid to the length of the attack.

### 2.2.2 Decay-related descriptors

Decay Spectral Flatness (6) is the ratio between the geometrical mean and the arithmetical mean (this gives an idea of the shape of the spectrum, if it's flat, the sound is more "white-noise"-like; if flatness is low, it will be more "musical"); Decay Spectral Centroid (7), which is the centre of gravity of the spectrum; Decay Strong Peak (8), intended to reveal whether the spectrum presents a very pronounced peak (the thinner and the higher the maximum of the spectrum is, the higher value takes this parameter); Decay Spectral Kurtosis (9), the 4<sup>th</sup> order central moment (it gives clues about the shape of the spectrum: "peaky" spectra have larger kurtosis than scattered or outlier-prone spectra.); Decay Zero-Crossing Rate (10); "Strong Decay" (11), a feature built from the non-linear combination of the energy and temporal centroid of a frame (a frame containing a temporal centroid near its left boundary and strong energy is said to have a "strong decay"); Decay Spectral Centroid Variance (12); Decay Zero-Crossing Rate Variance (13); and Decay Skewness (14), the 3<sup>rd</sup> order central moment (it gives indication about the shape of the spectrum in the sense that asymmetrical spectra tend to have large skewness values).

### 2.2.3 Relative energy descriptors

By dividing the spectrum of the decay part into 8 bands of frequency, the energy lying in them was calculated, and then the relative energy percent for each band was computed. These bands were basically chosen empirically, according to the observations of several spectra from relevant instruments. The boundaries were fixed after several trials in order to get significant results, and were the following: 40-70 Hz. (15), 70-110 Hz. (16), 130-145 Hz. (17), 160-190 Hz. (18), 300-400 Hz.(19), 5-7 KHz. (20), 7-10 KHz. (21), and 10-15 KHz. (22).

### 2.2.4 Mel-Frequency Cepstrum Coefficients

MFCC's have been usually used for speech processing applications, though they have shown usefulness in music applications too [25]. As they can be used as a compact representation of the spectral envelope, their variance was also recorded in order to keep some time-varying information. 13 MFCC's were computed over the whole signal, and their means and variances were used as descriptors. In order to interpret the selected sets of features in section 3, we will use the numeric ID's 23-35 for the MFCC means, and 36-48 for the MFCC variances.

## 2.3 Classification techniques

We have selected three different families of techniques to be compared<sup>1</sup>: instance-based algorithms, statistical modelling with linear functions, and decision tree building algorithms. The *K-Nearest Neighbors* (K-NN) technique is one of the most popular for instance-based learning and there are several papers on musical instrument sound classification using K-NN [26], [27], [28] [4]. As a novelty in this research context, we have also tested another instance-based algorithm called *K\** (pronounced "K-star"), which classifies novel examples by retrieving the nearest stored example using an entropy measure instead of an Euclidean distance. Systematic evaluations of this technique using standard test datasets [29] showed a significant improvement of performance over the traditional K-NN algorithm.

*Canonical discriminant analysis* is a statistical modelling technique that classifies new examples after deriving a set of orthogonal linear functions that partition the observation space into regions with the class centroids separated as far as possible, but keeping the variance of the classes as low as possible. It can be considered like an ANOVA (or MANOVA) that instead of continuous to-be-predicted variables uses discrete (categorical) variables. After a successful discriminant function analysis, "important" variables can be detected. Discriminant analysis has been successfully used by [30] for classification of wind and string instruments.

*C4.5* [31] is a decision tree technique that tries to focus on relevant features and ignores irrelevant ones for partitioning the original set of instances into subsets with a

---

<sup>1</sup> The discriminant analysis was run with SYSTAT (<http://www.spssscience.com/SYSTAT/>), and the rest of analyses with the WEKA environment ([www.cs.waikato.ac.nz/~ml/](http://www.cs.waikato.ac.nz/~ml/)).

strong majority of one of the classes. Decision trees, in general, have been pervasively used for different machine learning and classification tasks. Jensen and Arnspang [32] or Wiczorkowska [33] have used decision trees for musical instrument classification. An interesting variant of C4.5, that we have also tested, is PART (partial decision trees). It yields association rules between descriptors and classes by recursively selecting a class and finding a rule that "covers" as many instances as possible of it.

## 2.4 Cross-validation

For the forthcoming experiments the usual ten-fold procedure was followed: 10 subsets containing a 90% randomly selected sample of the sounds were selected for learning or building the models, and the remaining 10% was kept for testing them. Hit-rates presented below have been computed as the average value for the ten runs.

## 3. RESULTS

### 3.1 Selection of relevant descriptors

Two algorithm-independent methods for evaluating the relevance of the descriptors in the original set have been used: Correlation-based Feature Selection (hence CFS) and ReliefF. CFS evaluates subsets of attributes instead of evaluating individual attributes. A "merit" heuristic is computed for every possible subset, consisting of a ratio between how predictive a group of features is and how much redundancy or inter-correlation there is among those features [34]. Table 2 shows the CFS-selected features in the three different contexts of classification we are dealing with. Note that a reduction of more than fifty percent can be achieved in the most difficult case, and that the selected sets for basic level and for sub-category classification show an important overlap.

**Table 2. Features selected by the CFS method**

Super-category	[21, 4, 22]
Basic-level	[2, 4, 5, 6, 7, 9, 10, 14, 15, 16, 17, 18, 20, 21, 22, 26, 27, 30, 39]
Sub-category	[1, 2, 3, 4, 5, 6, 7, 9, 10, 14, 15, 16, 17, 18, 19, 20, 21, 22, 26, 30, 39]

ReliefF evaluates the worth of an attribute by repeatedly sampling an instance and considering the value of the given attribute for the nearest instance of the same and for the nearest different class [34]. Table 3 shows the ReliefF-selected features in the three different contexts. Note that the list is a ranked one –from most to least relevant– and that we have matched the cardinality of this list to the one yielded by the previous method, in order to facilitate their comparisons.

**Table 3. Features selected by the ReliefF method**

Super-category	[9, 14, 7]
Basic-level	[9, 14, 7, 19, 10, 17, 4, 25, 18, 6, 15, 21, 20, 16, 24, 26, 30, 31, 13]
Sub-category	[9, 14, 19, 7, 17, 10, 4, 25, 16, 15, 18, 6, 21, 20, 24, 30, 26, 31, 13, 28, 2]

Comparing the two methods it can be seen that all selected subsets for basic-level or for sub-category share more than 60% of features (but surprisingly they do not coincide at all when the target is the super-category). It is also evident that they include quite a heterogeneous selection of descriptors (some MFCC's, some energy bands, some temporal descriptors, some spectral descriptors...).

Contrasting with the previous “filtering” approaches, we also tested a “wrapping” approach for feature selection [35]. This means that features are selected in connection with a given classification technique which acts as a wrapper for the selection. Canonical Discriminant Analysis provides numerical indexes in order to decide about the relevance of a feature (but after analysis, not prior to it) as for example the F-to-remove value, or the descriptor's coefficients inside the canonical functions. For feature selection inside CDA it is usual to follow a stepwise (usually backwards) procedure. This strategy, however, only grants a locally optimal solution, so that an exhaustive (but sometimes impractical) search of all the combinations is recommended [36]. In our case, we have proceeded with a combination of backward stepwise plus some heuristic search. Table 4 shows the selected subsets, with the features ranked according the F-to-remove value (the most relevant first). A difference related to the filtering approaches is that with CDA the selected sets are usually larger. A large proportion of the selected features, otherwise, match those selected with the other methods.

**Table 4. Features selected after Canonical Discriminant Analyses**

Super-category	[4, 13, 19, 20, 37, 39]
Basic-level	[15, 9, 4, 20, 14, 2, 13, 26, 27, 3, 19, 8, 21, 39, 6, 11, 38]
Sub-category	[16, 15, 3, 9, 2, 17, 20, 13, 14, 19, 27, 26, 39, 7, 12, 10, 8, 37, 38, 4, 21, 22, 25, 33, 30, 29, 5, 24, 28, 45, 36, 34]

### 3.1 Classification results

We tested the three algorithms using the different subsets discussed in the previous section. Three different levels of classification were tested: super-category (plates versus membranes), basic-level (the five instruments) and sub-category (kick and snare plus some variations of the other three instruments: open and closed hihat, low, mid and high tom, crash and ride cymbal). Tables 5, 6 and 7 summarize the main results regarding hit rates for the three different classification schemes we have tested. Rows contain the different algorithms and columns contain the results using the different sets of features that were presented in the previous section. For the C4.5, the

number of leaves appears inside parentheses. For PART, the number of rules appears inside parentheses. The best method for each feature set has been indicated with bold type and the best overall result appears with grey background.

**Table 5. Super-category classification hit rates for the different techniques and feature selection methods**

	All features	CFS	ReliefF	CDA
K-NN (k=1)	<b>99.2</b>	97.9	93.7	96.7
K*	98.6	97.8	<b>94.8</b>	96.7
C4.5	97.2 (8)	98.6 (8)	<b>94.8 (12)</b>	95.1(14)
PART	98.4 (5)	98.2 (6)	94.4 (6)	95.1(9)
CDA	99.1	94.7	88.1	<b>99.3</b>

**Table 6. Basic-level classification hit rates for the different techniques and feature selection methods**

	All features	CFS	ReliefF	CDA
K-NN (k=1)	96.4	95	95.6	95.3
K*	<b>97</b>	<b>96.1</b>	<b>97.4</b>	<b>95.8</b>
C4.5	93 (20)	93.3(21)	92.2(23)	94.2(18)
PART	93.3 (12)	93.(11)	93.1(11)	93.6(12)
CDA	92	93	91	95.7

**Table 7. Sub-category classification hit rates for the different techniques and feature selection methods**

	All features	CFS	ReliefF	CDA
K-NN (k=1)	<b>89.9</b>	87.7	89.4	87.9
K*	<b>89.9</b>	<b>89.1</b>	<b>90.1</b>	<b>90.7</b>
C4.5	82.6 (40)	83 (38)	81 (45)	85(43)
PART	83.3 (24)	84.1(27)	81.9 (29)	84.3(27)
CDA	82.8	86	82	86.6

A clear interaction effect between feature selection strategy and algorithm family can be observed: for instance-based algorithms ReliefF provides the best results while for the decision-trees the best results have been obtained with CFS. In the case of decision trees, selecting features with CFS is good not only for improving hit-rates but also for getting more compact trees, (i.e. with a small number of leaves and therefore smaller in size). As expected, the CDA-selected features have yielded the best hit-rates for the CDA, but surprisingly they have also yielded the best hit-rates for most of the decision-trees.

It is interesting to compare the results obtained using feature selection with those obtained with the whole set of features. For the super-category classification it seems



that all the selection procedures have operated an excessive deletion and performance has degraded up to 4% when using a selected subset. Note however that in this classification test the best overall result (CDA features with CDA classification) outperforms any of the figures obtained with the whole subset. For the basic-level and sub-category tests, the reduction of features degrades the performance of instance-based methods (but less than 1%), whereas it improves the performance of the rest.

After comparing families of algorithms it is clear that differences between them increase as the task difficulty increases. It is also evident that the best performance is usually found in instance-based ones (and specifically K\* yields slightly better results than a simple K-NN), whereas tree-based yield the worst figures and CDA lies in between. Although decision trees do not provide the best overall performance, they have an inherent advantage over instance-based: expressing relationships between features and classes in terms of conditional rules. Table 8 exemplifies the type of rules that we get after PART derivation.

**Table 8. Some of the PART rules for classification at the "basic-level". Correctly and wrongly classified instances are shown inside parentheses. We have left out some less general rules for clarity**

SKEWNESS > 4.619122 AND B40HZ70HZ > 7.784892 AND MFCC3 <= 1.213368: Kick (105.0/0.0)	SPECCENTROID > 11.491498 AND B1015KHZ > 0.791702: HH (100.0/2.0)
KURTOSIS > 26.140138 AND TEMPORALCE <= 0.361035 AND ATTZCR > 1.478743: Tom (103.0/0.0)	SKEWNESS <= 4.485531 AND B160HZ190HZ <= 5.446338 AND MFCC3VAR > 0.212043 AND MFCC4 > -0.435871: Cymbal (110.0/3.0)
B710KHZ <= 0.948147 AND KURTOSIS <= 26.140138 AND ATTZCR <= 22.661397: Snare (133.0/0.0)	

Regarding CDA, an examination of the canonical scores plots provides some graphical hints about the performance of the four canonical discriminant functions needed for the basic-level case: the first one separates toms+kicks from hihats+cymbals, the second one separates the snare from the rest, the third one separates cymbals from hihats, and the fourth one separates toms from kicks. It should be noted that in the other cases it is more difficult to assign them a clear role.

Inspecting the confusion matrix for the instrument test, most of the errors consist in confusing cymbals with hihat, and tom with kick (and their inverse confusions, though with a lesser incidence). For the sub-instrument test, 60% of the misclassifications appear to be intra-category (i.e. between crash and ride, between open and closed hihat, etc.), and they are evenly distributed.

#### 4. DISCUSSION

We have achieved very high hit rates for the automatic classification of standard drum sounds into three different classification schemes. The fact that, in spite of using three

very different classification techniques, we have obtained quite similar results could mean that the task is quite an easy one. It is true that the number of categories we have used has been small even for the most complex classification scheme. But it should also be noted that there are some categories that, at least from a purely perceptual point of view, do not seem to be easily separated (for example, low-toms from some kicks, or some snares from mid-toms or from some crash cymbals). Therefore, a contrasting additional interpretation for this good performance is to consider that our initial selection of descriptors was good. This statement gets support by the fact that the all-feature results are not much worse than results after feature selection. In the case of having a bad initial set, those bad features would have contributed to worsen the performance. As it has not been the case, we can conclude that from a good set of initial features, some near-optimal sets have been identified with the help of filtering or wrapping techniques. Most of the best features found can be considered as spectral descriptors: skewness, kurtosis, centroid, MFCC's. We included a very limited number of temporal descriptors, but, as expected, apart from ZCR, they do not seem to be needed for precise instrument classification.

In the section of improvements for subsequent research we may list the following: (1) A more systematic approach to description in terms of energy bands (for example, using Bark measures); (2) Evaluation of whole-sound descriptors against attack-decay decomposed descriptors (i.e. the ZCR); (3) Non-linear scaling of some feature dimensions; (4) Justified deletion of some observations (after analyzing the models, it seems that some outliers that contribute to the increment of the confusion rates should be considered as "bad" examples for the model because of audio quality or wrong class adscription).

## **5. CONCLUSIONS**

In this study, we have performed a systematic study of the classification of standard drum sounds. After careful selection of descriptors and its refinement with different techniques, we have achieved very high hit-rates in three different classification tasks: super-category, basic-level category, and sub-category. In general, the most relevant descriptors for them seem to be ZCR, kurtosis, skewness, centroid, relative energy in specific bands, and some low-order MFCC's. Performance measures classification techniques have not yielded dramatic differences between classification techniques and therefore selecting one or another is clearly an application-dependent issue. We believe, though, that relevant performance differences will arise when more classes are included in the test, as we have planned for a forthcoming study. Regarding classification of mixtures of sounds, even if it is not yet clear if the present results will be useful, we have gathered interesting and relevant data in order to characterize different classes of drum sounds.

## **6. ACKNOWLEDGMENTS**

The research reported in this paper has been partially funded by the EU-IST project CUIDADO.

## REFERENCES

- [1] Zhang, T. and Jay Kuo, C.-C.: Classification and retrieval of sound effects in audiovisual data management. In Proceedings of 33rd Asilomar Conference on Signals, Systems, and Computers (1991)
- [2] Casey, M.A.: MPEG-7 sound recognition tools. IEEE Transactions on Circuits and Systems for Video Technology, 11, (2001) 37-747
- [3] Herrera, P., Amatriain, X., Batlle, E., Serra, X.: A critical review of automatic musical instrument classification. In Byrd, D, Downie, J.S., and Crawford, T (Eds.), Recent Research in Music Information Retrieval: Audio, MIDI, and Score Kluwer Academic Press, in preparation.
- [4] Kaminskyj, I.: Multi-feature Musical Instrument Sound Classifier. In Proceedings of Australasian Computer Music Conference (2001)
- [5] Schloss, W.A.: On the automatic transcription of percussive music -from acoustic signal to high-level analysis. STAN-M-27. Stanford, CA, CCRMA, Department of Music, Stanford University (1985)
- [6] Bilmes, J.: Timing is the essence: Perceptual and computational techniques for representing, learning and reproducing expressive timing in percussive rhythm. MSc, Thesis. Massachusetts Institute of Technology, Media Laboratory. Cambridge, MA. (1993)
- [7] McDonald, S. and Tsang, C.P.: Percussive sound identification using spectral centre trajectories. In Proceedings of 1997 Postgraduate Research Conference (1997)
- [8] Sillanpää, J.: Drum stroke recognition. Tampere University of Technology. Tampere, Finland (2000)
- [9] Sillanpää, J., Klapuri, A., Seppänen, J., and Virtanen, T.: Recognition of acoustic noise mixtures by combined bottom-up and top-down approach. In Proceedings of European Signal Processing Conference, EUSIPCO-2000 (2000)
- [10] Goto, M., Muraoka, Y.: A sound source separation system for percussion instruments. Transactions of the Institute of Electronics, Information and Communication Engineers D-II, J77, 901-911 (1994)
- [11] Goto, M. and Muraoka, Y.: A real-time beat tracking system for audio signals. In Proceedings of International Computer Music Conference, 171-174 (1995)
- [12] Gouyon, F. and Herrera, P.: Exploration of techniques for automatic labeling of audio drum tracks' instruments. In Proceedings of MOSART: Workshop on Current Directions in Computer Music (2001)
- [13] Miller, J.R., Carterette, E.C.: Perceptual space for musical structures. Journal of the Acoustical Society of America, 58, 711-720 (1975)
- [14] Grey, J.M.: Multidimensional perceptual scaling of musical timbres. Journal of the Acoustical Society of America, 61, 1270-1277 (1977)
- [15] McAdams, S., Winsberg, S., de Soete, G., and Krimphoff, J.: Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject classes. Psychological Research, 58, 177-192 (1995)
- [16] Toiviainen, P., Kaipainen, M., and Louhivuori, J.: Musical timbre: Similarity ratings correlate with computational feature space distances. Journal of New Music Research, 282-298 (1995)
- [17] Lakatos, S.: A common perceptual space for harmonic and percussive timbres. Perception and Psychophysics, 62, 1426-1439 (2000)

- [18] McAdams, S., Winsberg, S.: A meta-analysis of timbre space. I: Multidimensional scaling of group data with common dimensions, specificities, and latent subject classes (2002)
- [19] Peeters, G., McAdams, S., and Herrera, P.: Instrument sound description in the context of MPEG-7. In Proceedings of Proceedings of the 2000 International Computer Music Conference (2000)
- [20] Scavone, G., Lakatos, S., Cook, P., and Harbke, C.: Perceptual spaces for sound effects obtained with an interactive similarity rating program. In Proceedings of International Symposium on Musical Acoustics (2001)
- [21] Laroche, J., Meillier, J.-L.: Multichannel excitation/filter modeling of percussive sounds with application to the piano. *IEEE Transactions on Speech and Audio Processing*, 2, (1994) 329-344
- [22] Repp, B.H.: The sound of two hands clapping: An exploratory study. *Journal of the Acoustical Society of America*, 81, (1993) 1100-1109
- [23] Freed, A.: Auditory correlates of perceived mallet hardness for a set of recorded percussive events. *Journal of the Acoustical Society of America*, 87, (1990) 311-322
- [24] Klatzky, R.L., Pai, D.K., and Krotkov, E.P.: Perception of material from contact sounds. *Presence: Teleoperators and Virtual Environments*, 9, (2000) 399-410
- [25] Logan, B.: Mel Frequency Cepstral Coefficients for Music Modeling. In Proceedings of International Symposium on Music Information Retrieval, ISMIR-2000. Plymouth, MA, (2000)
- [26] Martin, K.D. and Kim, Y.E.: Musical instrument identification: A pattern-recognition approach. In Proceedings of Proceedings of the 136th meeting of the Acoustical Society of America. (1998)
- [27] Fujinaga, I. and MacMillan, K.: Realtime recognition of orchestral instruments. In Proceedings of the 2000 International Computer Music Conference, (2000) 141-143
- [28] Eronen, A.: Comparison of features for musical instrument recognition. In Proceedings of 2001 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'01) (2001)
- [29] Cleary, J.G. and Trigg, L.E.: K\*: An instance-based learner using an entropic distance measure. In Proceedings of International Conference on Machine Learning, (1995) 108-114
- [30] Agostini, G., Longari, M., and Pollastri, E.: Musical instrument timbres classification with spectral features. In Proceedings of IEEE Multimedia Signal Processing Conference (2001)
- [31] Quinlan, J.R.: C4.5: Programs for machine learning. Morgan Kaufmann. San Mateo, CA, (1993)
- [32] Jensen, K. and Arnspang, J.: Binary decision tree classification of musical sounds. In Proceedings of the 1999 International Computer Music Conference. (1999)
- [33] Wiczorkowska, A.: Classification of musical instrument sounds using decision trees. In Proceedings of the 8th International Symposium on Sound Engineering and Mastering, ISSEM'99, (1999) 225-230
- [34] Hall, M.A.: Correlation-based feature selection for discrete and numeric class machine learning. In Proceedings of Seventeenth International Conference on Machine Learning (2000)
- [35] Blum, A., Langley, P.: Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97, (1997) 245-271
- [36] Huberty, C.J.: Applied discriminant analysis. John Wiley. New York (1994)

**Herrera, P.,** Peeters, G., Dubnov, S. (2003). "Automatic Classification of Musical Instrument Sounds". Journal of New Music Research. 32(1), pp. 3-21.

DOI: <https://doi.org/10.1076/jnmr.32.1.3.16798>

ISSN: 0929-8215

Online ISSN 1744-5027



---

## Automatic Classification of Musical Instrument Sounds

---

Perfecto Herrera-Boyer<sup>1</sup>, Geoffroy Peeters<sup>2</sup> and Shlomo Dubnov<sup>3</sup>

<sup>1</sup>MTG-IUA, Universitat Pompeu Fabra, Barcelona, Spain, <sup>2</sup>IRCAM, Paris, France and <sup>3</sup>The Hebrew University, Jerusalem, Israel

---

### Abstract

We present an exhaustive review of research on automatic classification of sounds from musical instruments. Two different but complementary approaches are examined, the perceptual approach and the taxonomic approach. The former is targeted to derive perceptual similarity functions in order to use them for timbre clustering and for searching and retrieving sounds by timbral similarity. The latter is targeted to derive indexes for labeling sounds after culture- or user-biased taxonomies. We review the relevant features that have been used in the two areas and then we present and discuss different techniques for similarity-based clustering of sounds and for classification into pre-defined instrumental categories.

### 1. Introduction

The need for automatic classification of sounds arises in different contexts: biology (e.g., for identifying animals belonging to a given species, or for cataloguing communicative resources) (Fristrup & Watkins, 1995; Mills, 1995; Potter, Mellinger, & Clark, 1994) medical diagnosis (e.g., for detecting abnormal conditions of vital organs) (Shiyong, Zehan, Fei, Li, & Shouzong, 1998; Buller & Lutman, 1998; Schön, Puppe, & Manteuffel, 2001) surveillance (e.g., for recognizing machine-failure conditions) (McLaughling, Owsley, & Atlas, 1997) military operations (e.g., for detecting an enemy engine approaching or for weapon identification) (Gorman & Sejnowski, 1988; Antonic & Zagar, 2000; Dubnov & Tishby, 1997) and multimedia content description (e.g., for helping video scene classification or object detection) (Liu, Wang, & Chen, 1998; Pfeiffer, Lienhart, & Effelsberg, 1998). Speech, sound effects, and music are the three main sonic categories that are combined in multimedia databases. Describing

multimedia sound therefore means describing each one of those categories. In the case of speech, the main description concerns speaker identification and speech transcription. Describing sound effects means determining the apparent sound source, or clustering similar sounds even though they have been generated by different sources. In the case of music, description calls for deriving indexes in order to locate melodic patterns, harmonic or rhythmic structures, musical instrument sets, usage of expressivity resources, etc. As we are not concerned here with discrimination between speech, music and sound effects, we recommend interested readers consult the work by Zhang and Kuo (1998b; 1999a).

Provided that we are interested in a music-only stream of audio data, one of the most important description problems is the correct identification of the musical instruments present in the stream. This is a very difficult task that is far from being solved. The practical utility for musical instrument classification is twofold:

- First, to provide labels for monophonic recordings, for “sound samples” inside sample libraries, or for new patches created with a given synthesizer;
- Second, to provide indexes for locating the main instruments that are included in a musical mixture (for example, one might want to locate a saxophone “solo” in the middle of a song);

The first problem is easier to solve than the second one, and it seems clearly solvable given the current state of the art, as we will see later in this paper. The second is tougher, and it is not clear if research done on solving the first one may help.

Common sense dictates that a reasonable approach to the second problem would be the initial separation of the sounds corresponding to the different sound sources, followed by the

---

Accepted: 9 May, 2002

*Correspondence:* Perfecto Herrera-Boyer, Institut de l'Audiovisual, Universitat Pompeu Fabra, Pg. Circumval.lació 8, 08003 Barcelona, Spain. Tel.: +34 93 5422 806, Fax: +34 9 25242206, E-mail: perfecto.herrera@iua.upf.es

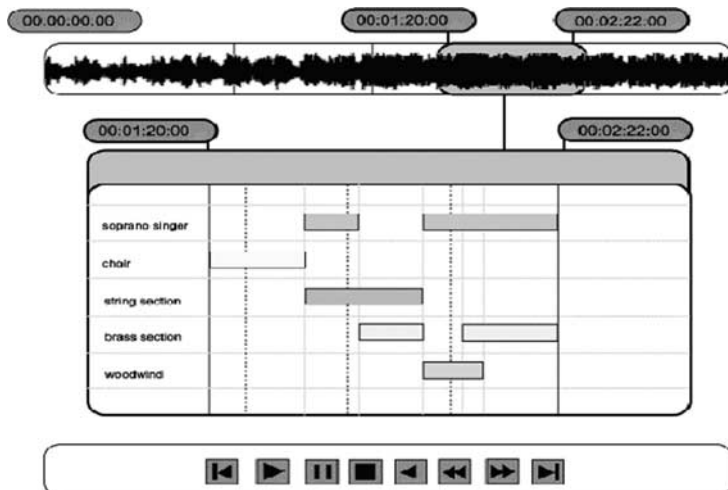


Fig. 1. An imaginary instrument browser adapted from Smoliar and Wilcox (Smoliar & Wilcox, 1997).

segmentation<sup>1</sup> and classification<sup>2</sup> on those separated tracks. Techniques for source separation cannot yet provide satisfactory solutions although some promising approaches have been developed (Casey & Westner, 2001; Ellis, 1996; Bell & Sejnowski, 1995; Varga & Moore, 1990). As a consequence, research on classification has concentrated on working with isolated sounds under the assumption that separation and segmentation have been previously performed. This implies the use of a sound sample collection (usually isolated notes) consisting of different instrument families and classes. The general classification procedure can be described as follows:

- Lists of features are selected to describe the samples.
- Values for these features are computed.
- A learning algorithm that uses the selected features to discriminate between instrument families or classes is applied.

<sup>1</sup>Segmentation can be defined as the process of breaking up an audio stream into temporal segments by means of applying a boundary detection criterion as, for example, texture, note, instrument, rhythm pattern, overall structure, etc. The same audio stream can be segmented in different ways by recurrently applying different criteria.

<sup>2</sup>Once an audio stream has been segmented, labels have to be attached to the segments. Two different families of algorithms can be used for learning labels: in the case we know in advance the labels to be used, *pattern recognition*, *discrimination*, or *supervised learning* techniques are the logical choice; when we do not know beforehand the labels and they will have to be inferred from the data, then the right choice is some *unsupervised learning* or *clustering* technique. See {Michie, Spiegelhalter, et al., 1994 109 /id} for more details.

- The performance of the learning procedure is evaluated by classifying new sound samples (cross-validation).

Note that there is a very important tradeoff in endorsing this isolated-notes strategy: we gain simplicity and tractability, but we lose contextual and time-dependent cues that can be exploited as relevant features for classifying musical sounds in complex mixtures. It is also important to note that the implicit assumption that solutions for isolated sounds can be extrapolated to complex mixtures should not be taken for granted, as we will discuss in the final section. Another implicit assumption that should not be taken for granted is that the arbitrary taxonomy that we use is optimal or, at least, good for the task (see Kartomi, 1990, for issues regarding arbitrary taxonomies of musical instruments).

An alternative approach to the whole problem is to shift focus from the traditional *transcription* concern to that of *description* or *understanding* (Scheirer, 2000). This is what some Computational Auditory Scene Analysis systems have addressed (Ellis, 1996; Kashino & Murase, 1997a). We will return to this distinction later but for the moment a clarifying practical example of this different focus can be provided with an “instrument browser” as the one depicted in Figure 1. In order to develop this kind of application, we only need to detect the instrument *boundaries*. The boundaries can surround individual instruments or classes of instruments (Aucouturier & Sandler, 2001). For example, note how the “soprano singer” instrument has been drawn separately whereas the other instruments are grouped into classes. In Figure 1, the string section subsumes the phrases played by violins, violas and cellos. The goal of this approach is not to separate into distinct tracks each of the instrumental voices but, rather, to label their locations within the context of the



musical work. Thus, the user, when clicking on one of the labels would not hear an isolated instrument; instead, the user would be taken to part of the piece where the desired instrument or instrument family can be clearly heard. Manipulating the source file to bring to the foreground the selected instrument(s) is a possible enhancement of this boundary-based approach. In order to develop that kind of application we *only* need to detect the instrument boundaries.

A very different type of classification arises when our target is not an instrument class but a cluster of sounds that can be judged to be perceptually similar. In that case, classification does not rely on culturally shared labels but on timbre similarity measures and distance functions derived from psychoacoustical studies (Grey, 1977; Krumhansl, 1989; McAdams, Winsberg, de Soete, & Krimphoff, 1995; Lakatos, 2000). This type of perceptual classification or clustering is addressed to provide indexes for retrieving sounds by similarity, using a query by example strategy.

In the following sections we review the different features (perceptual-based or taxonomic-based) that have been used for musical sound classification, and then the techniques that have been tested for classification and clustering of isolated sounds. We have purposely refrained from writing mathematical formulae in order to facilitate the basic understanding to casual readers. It is our hope that the comprehensive list of references at the end of the chapter will compensate this lack, and will help in finding the complementary technical information that a thorough comprehension requires.

## 2. Perceptual description versus taxonomic classification

Perceptual description departs from taxonomic classification in that it tries to find features that explain human perception of sounds, while the latter is interested in assigning to sounds some label from a previously established taxonomy (family of musical instruments, instruments names, sound effects category . . .). Therefore, the latter may be considered deterministic while the former is derived from experimental results using human subjects or artificial systems that simulate some of their perceptual processes.

Perception of sounds has been studied systematically since Helmholtz. It is now well accepted that sounds can be described in terms of their pitch, loudness, subjective duration, and something called “timbre.” According to the ANSI definition (American National Standards Institute, 1973) timbre refers to the features that allow one to distinguish two sounds that are equal in pitch, loudness, and subjective duration. The underlying perceptual mechanisms are rather complex but they involve taking into account several perceptual dimensions at the same time in a possibly complex way. Timbre is thus a multi-dimensional sensation that relies among others, on spectral envelope, temporal envelope, and on variations of each of them. In order to understand better what the timbre feature refers to, numerous experiments have

been performed (Plomp, 1970; Plomp, 1976; Wedin & Goude, 1972; Wessel, 1979; Grey, 1977; Krumhansl, 1989; McAdams, Winsberg, de Soete, & Krimphoff, 1995; Lakatos, 2000). In all of these experiments, people were asked for a dis-similarity judgment on pairs of sounds. Multidimensional Scaling (MDS) analysis<sup>3</sup> was used to process the judgments, and to represent the sound stimuli in a low-dimensional space revealing the underlying attributes used by listeners when making the judgments. Researchers often refer to this low-dimensional representation as a “Timbre Space” (see Fig. 2).

Grey (1977) performed one of the first experiments under this paradigm. Using 16 instrument sounds from the orchestra (string and wind instruments) he derived from MDS a timbre space with 3 dimensions corresponding to the main perceptual axes. A qualitative description of these axes allowed him to assign one dimension to the spectral energy distribution, another to the amount of synchronicity of the transients and amount of spectral fluctuation, and the last one to the temporal attribute of the beginning of the sound.

Wessel’s experiments (Wessel, 1979) used the 16 sounds from Grey (1977) plus 8 hybrid sounds (in order to use non-existing sounds that avoided the class recognition effects and also for getting intermediate “timbral steps” between sounds). This research yielded a 2-dimensional space with one dimension assigned to the “brightness” of the sustained part of the sound, and the other to the steepness of the attack and the offset between the beginnings of the high frequency harmonics to the low frequency ones.

Krumhansl (1989) used 21 FM-synthesis sounds from Wessel, Bristow, and Settel (1987) mainly sustained harmonic sounds. She found the same results as Grey, but assigned the third dimension to something called “spectral flux” that was supposed to be related to the variations of the spectral content along time. McAdams et al. (1995) also used these 21 FM-synthesis sounds in a new experiment and tested a new MDS technique that estimates the latent classes of subjects, instrument specificity values, and separate weights for each class. Compared to Krumhansl’s results, they confirmed the assignment of one dimension to the attack-time, another to the spectral centroid, but they did not confirm the “spectral flux” for the last dimension.

Lakatos’ experiment (2000) used 36 natural sounds from the McGill University sound library, both wind and string (17) and percussive (18) sounds. The goal of this experiment was to extend the timbre space to percussive and mixed percussive/sustained sounds. This yields a two dimensional space and a three dimensional space. The conclusion of the experiment is that, except for spectral centroid and rise time, additional perceptual dimensions exist but their precise

<sup>3</sup>Multidimensional Scaling is a technique for discovering the number of underlying dimensions appropriate for a set of multidimensional data and for locating the observations in a low-dimensional space (Wish & Carroll, 1982).

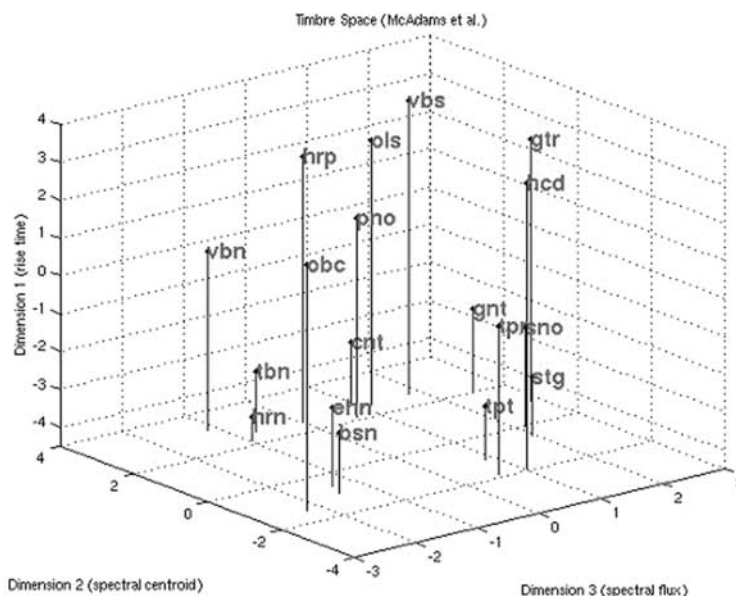


Fig. 2. Timbre Space coming from McAdams et al. (1995) experiment. It was derived from dissimilarity ratings on 18 timbres by 88 subjects with specificities and five latent subject classes. Acoustic correlates of the three dimensions: rise time, spectral centroid, spectral flux (reproduced here with permission).

acoustic correlates are context dependent and therefore less prominent.

An interesting practical application of this similarity-based research is that of setting up a given orchestration with some set of reference sound samples and then substituting some of them without radically changing the orchestration. Practical reasons for doing query-by-similarity of sound samples could include performance rights or copyrights issues, sample format compatibility, etc. Working examples of the timbre similarity approach are, for example, the *Soundfisher* system developed by Musclefish<sup>4</sup>, and the *Studio On Line* developed by IRCAM<sup>5</sup>. *Soundfisher*, recently incorporated as a plug-in into a commercial video-logger called Virage, is designed to perform the classification, indexing and search of sounds in general, though it can be used in a music context. The initial versions of *Soundfisher*<sup>6</sup> (Keislar, Blum, Wheaton, & Wold, 1995) did not yield an explicit class decision but, rather, generated a list of mathematically similar sounds. Some kind of class decision procedure, however, seems to have been recently implemented (Keislar, Blum, Wheaton, & Wold, 1999). The *Soundfisher* system implicitly implements the assumption that what is mathematically similar can be also considered perceptually

similar; in other words, that the computed features accurately represent perceptual dimensions, an assumption that contradicts most empirical studies. In contrast, *Studio On Line* computes similarity by using features that have been extracted under the paradigm of the above-cited perceptual similarity psychoacoustical experiments. The interested reader can find in Peeters, McAdams, and Herrera (2000) a recent validation of the psychoacoustical approach in the context of MPEG-7.

### 3. Relevant features for classification

#### 3.1 Types of features

The term *feature* denotes a quantity or a quality<sup>7</sup> describing an object of the world. In the realm of signal processing and pattern recognition, objects are usually described by using vectors or lists of features. Features are also known as *attributes* or *descriptors*. Audio signal features are usually computed directly from the signal, or from the output yielded by transformations such as the Fast Fourier Transform or the Wavelet Transform. These audio signal features are usually computed every few milliseconds, for a very short segment of audio samples, in order to grasp their micro-temporal evolution. Macro-temporal evolution features can also be com-

<sup>4</sup> <http://www.musclefish.com>

<sup>5</sup> <http://www.ircam.fr/produits/technologies/sol/index-e.html>

<sup>6</sup> <http://www.soundfisher.com>

<sup>7</sup> In this paper we will only consider the quantitative approach.

puted by using a longer segment of samples (e.g., attack time, vibrato rate . . .) or by summarizing micro-temporal values (e.g., averages, variances . . .).

A systematic taxonomy of features is outside the scope of this paper; nevertheless we could distinguish features at least according to four points of view:

1. The steadiness or dynamicity of the feature, i.e., the fact that the features represent a value extracted from the signal at a given time, or a parameter from a model of the signal behavior along time (mean, standard deviation, derivative or Markov model of a parameter);
2. The time extent of the description provided by the features: some description applies to only part of the object (e.g., description of the attack of the sound) whereas other apply to the whole signal (e.g., loudness);
3. The “abstractness”, i.e., what the feature represents (e.g., cepstrum and linear prediction are two different representation and extraction techniques for representing spectral envelope, but probably the former one can be considered as more abstract than the latter);
4. The extraction process of the feature. According to this point of view, we could further distinguish:
  - Features that are directly computed on the waveform data as, for example, zero-crossing rate (the rate that the waveform changes from positive to negative values);
  - Features that are extracted after performing a transform of the signal (FFT, wavelet . . .) as, for example, spectral centroid (the “gravity center” of the spectrum);
  - Features that relate to a signal model, as for example the sinusoidal model or the source/filter model;
  - Features that try to mimic the output of the ear system (bark or erb bank filter output).

### 3.2 Relevant features for perceptual classification

For each of the “timbre” experiments, people have tried to *qualify* the dimensions of these timbre spaces, the perceptual axes, in terms of “brightness,” “attack,” etc. Only recently attempts have been made to *quantitatively* describe these perceptual axes, i.e., relate the perceptual axes to variables or descriptors directly derived from the signal (Grey, 1978; Krimphoff, McAdams, & Winsberg, 1994; Misdariis, Smith, Pressnitzer, Susini, & McAdams, 1998).

This quantitative description is done by finding the signal features that best explain the dis-similarity judgment. This is usually done using regression or multiple-regression between feature values and sound positions in the “timbre” space, and keeping only the features that yield the largest correlation. This makes the perceptual description framework different from taxonomic classification, since in the latter we’re not looking at features that “best explain” but at features that allow to “best discriminate” (between the considered classes).

In the Grey and Gordon (1978) experiment, only one dimension correlated significantly with a perceptual dimen-

sion of their “timbre” space: the spectral centroid. Krimphoff et al. (1994) worked with Krumhansl’s space (1989) trying to find the quantitative parameters corresponding to its qualitative features and found, as Grey did, significant correlations with the spectral centroid, but also with the logarithm of the attack time and what they called the “spectral irregularity,” which is the average departure of the spectral harmonic amplitudes from a global spectral envelope. Krumhansl (1989) had labelled this dimension as “spectral flux.” Misdariis, Smith, Pressnitzer, Susini, and McAdams (1998) combined results coming from the Krumhansl (1989) and McAdams et al. (1995) experiments. They found the same features as Krimphoff did plus a new one that explained one dimension of McAdams et al. (1995) experiment: spectral flux defined here as the average of the correlation between amplitude spectra in adjacent time windows.

Peeters et al. (2000) considered also the two above-cited experiments by Krumhansl and McAdams et al., called here “sustained harmonic sound space” as opposed to the “percussive sound space” coming from Lakatos (2000) experiment. Two methods were used for the selection of the features, a “position” method, which tries to explain from the feature values the position of the sound in the timbre space, and a “distance” method, which tries to explain directly the perceived distance between sounds from a difference of feature values. From this study the following features, now part of the MPEG-7 standard, have been derived to describe the perceived similarity. For the “harmonic sustained sounds”: log-attack time, harmonic spectral centroid, harmonic spectral spread (the extent of the spectrum’s energy around the spectral centroid) harmonic spectral variation (the amount of variation of the spectrum energy distribution along time) and harmonic spectral deviation (the deviation of the spectrum harmonic from a global envelope). For the “percussive sounds”: log-attack time, temporal centroid (the temporal centre of gravity of the signal energy) and spectral centroid (the centre of gravity of the power spectrum of the whole sound).

Another approach is the one taken by the company Muscle Fish in the development of the *Soundfisher* system (Wold, Blum, Keislar, & Wheaton, 1966). In this case the selected features are not derived from experiments but they constitute a set that is similar to the one discussed above: loudness (rms value in dB) pitch, brightness (spectral centroid) bandwidth (spread of the spectrum around the spectral centroid) harmonicity (amount of energy of the signal explained by a periodic signal model) . . . In order to capture the temporal trend of the features, it is proposed to store their average, variance and auto-correlation values along time.

### 3.3 Relevant features for taxonomic classification

Mel-Frequency Cepstrum Coefficients (hence MFCCs) are features that have proved useful for such speech processing tasks as, for example, speaker identification and speaker recognition (Rabiner & Juang, 1993). MFCCs are computed

by taking the log of the power spectrum of a windowed signal, then non-linearly mapping the spectrum coefficients in a perceptually-oriented way (inspired by the Mel scale). This mapping is intended to emphasize perceptually meaningful frequencies. The Mel-weighted log-spectrum is then compacted into cepstral coefficients through the use of a discrete cosine transform. This transformation reduces the dimensionality of the representation without losing information (typically, the power spectrum may contain 256 values, whereas the MFCCs are usually less than 15). MFCCs provide a rather compact representation of the spectral envelope and are probably more musically meaningful than other common representations like Linear Predictive Coding coefficients or curve-fitting approximations to spectrum. Despite these strengths, MFCCs by themselves can only convey information about static behavior and, as a consequence, temporal dynamics cannot be considered. Another important drawback is that MFCCs do not have an obvious direct interpretation, though they seem to be related (in an abstract way) with the resonances of instruments. Despite these shortcomings Marques (1999) used MFCCs in a broad series of classification studies. Eronen and Klapuri (2000) used Cepstral Coefficients (without the Mel scaling) and combined these features with a long list (up to 43) of complementary descriptors. Their list included, among others, centroid, rise and decay time, FM/AM rate and width, fundamental frequency and fundamental-variation-related features for onset and for the remainder of the note. In a more recent study, using a very large set of features (Eronen, 2001), the most important ones seemed to be the MFCCs, their standard deviations, and their deltas (differences between contiguous frames) the spectral centroid and related features, onset duration, and crest factor (specially for instrument family discrimination). There are ways, however, for adding temporal information into a MFCCS classification schema. For example, Cosi, De Poli, and Prandoni (1994) created a Kohonen Feature Map<sup>8</sup> (Kohonen, 1995) using both note durations and the feature coefficients. The network then clustered and mapped the right temporal sequence into a bi-dimensional space. As a result, sounds were clustered in a human perceptual-like way (i.e., not into taxonomic classes but into timbrically similar conglomerates). Brown (1999) used cepstral coefficients from constant-Q transforms instead of taking them after FFT-transforms; she also clustered feature vectors in a way that the resulting clusters seemed to be coding some temporal dynamics.

One of the most commonly used descriptors for musical, as well as non-musical, sound classification is energy. In

Kaminskyj and Materka (1995) Root Mean Square (RMS) energy was used for classifying 4 different types of instruments with a neural network. In an additional, but apparently unfinished extension of this work (Kaminskyj & Voumard, 1996) the authors also included brightness, spectral onset asynchrony, harmonicity and MFCCs. In a more recent and comprehensive work (Kaminskyj, 2001) the main author used the RMS envelope, the Constant-Q frequency spectrum, and a set of spectral features derived from Principal Component Analysis (PCA from now on). PCA is commonly used to reduce dimensionality of complex data sets with a minimum loss of information. In PCA data is projected into abstract dimensions that are contributed with different – but partially related – variables. Then PCA calculates which projections, amongst all possible, are the best for representing the structure of data. The projections are chosen so that the maximum variability of the data is represented using the smallest number of dimensions. In this specific research, the 177 spectral bins of the Constant-Q were reduced, after PCA, to 53 “abstract” features without any significant loss in discriminative power.

Martin and Kim (1998) exemplified the idea of testing very long lists of features and then selecting only those shown to be most relevant for performing classifications. Martin and Kim worked with log-lag correlograms to better approximate the way our hearing system processes sonic information. They examined 31 features to classify a corpus of 14 orchestral wind and string instruments. They found the following features to be the most useful: vibrato and tremolo strength and frequency, onset harmonic skew (i.e., the time difference of the harmonics to arise in the attack portion) centroid related measures (e.g., average, variance, ratio along note segments, modulation) onset duration, and select pitch related measures (e.g., value, variance). The authors noted that the features they studied exhibited non-uniform influences, that is, some features were better at classifying some instruments and instrument families and not others. In other words, features could be both relevant and non-relevant depending on the context. The influence of non-relevant features degraded the classification success rates between 7% and 14%. This degradation is an important theoretical issue (Blum & Langley, 1997) that unfortunately has been overlooked by the majority of studies we have reviewed. It should be noted that there are some classification techniques that also provide some indication about the relevance of the involved features. This is the case with Discriminant Analysis (see section 4.2.3). Using this technique “backward” deletion and “forward” addition of features can be used in order to settle into a good (though sometimes suboptimal) set. Agostini, Longari, and Pollastri (2001) have used this method for reducing their original set of eighteen features to the eight ones that best separate the groups. The best features were: inharmonicity mean, centroid mean and standard deviation, harmonicity energy mean, zero-crossing rate, bandwidth mean and standard deviation, and standard deviation of harmonic skewness.

<sup>8</sup> A Kohonen or Self Organized Feature Map is a type of neural network that uses a single layer of interconnected units in order to learn a compact representation (i.e., with reduced features) of similar instances. It is very useful to cluster objects or instances that share some type of similarity because it preserves the inner space topology.

Spectral flatness is a feature that has been recently used in the context of MPEG-7 (Herre, Allamanche, & Hellmuth, 2001) for robust retrieval of song archives. It is a “new-comer” in musical instrument classification but can be quite useful because it indicates how flat (i.e., “white-noisy”) the spectrum of a sound is. Our current work indicates that it can also be a good descriptor for percussive sound classification (Herrera, Yeterian, & Gouyon, 2002).

Jensen and Arnspang (1999) used amplitude, brightness, tristimulus, amplitude of odd partials, irregularity of spectral envelope, shimmer and jitter measures, and inharmonicity, for studying the classification of 1500 sounds from 7 instruments. Jensen (1999) using PCA, had earlier identified these features as the most relevant from an initial set of 20 and indicated 3 relevant dimensions that could summarize the most important features. He labeled these, in decreasing order of importance, “spectral envelope,” “(temporal) envelope,” and “noise.” Kashino and Murase (1997b) applied PCA to the instrument classification problem: 41 features were reduced to 11. PCA, in the context of sound classification, can be also found in the works of Sandell and Martens (1995) and Rochebois and Charbonneau, (1997). Less compact representations for temporal or spectral envelopes can be found in Fragoulis, Avaritsiotis, and Papaodysseus (1999) who used the slope of the first five partials, the time delay in the onset of these partials, and the high-frequency energy. Cemgil and Grgen (1997) also used a set of harmonics (the first twelve) as discriminative features in their neural networks study.

Apart from PCA, another useful method for reducing the dimensions of the feature selection problem is the application of Genetic Algorithms (GAs). GAs are modeled on the processes that drive the evolution of gene populations (e.g., crossover, mutation, evaluation of fitness, and selection of the *best adapted*). GAs have a property called *implicit search*, which means that near-optimal combinations of genes can be found without explicitly evaluating all possible combinations. GAs have been used in other musical contexts (e.g., sound synthesis and music composition) but the only known application to sound classification has been that of Fujinaga, Moore, and Sullivan (1998) where GAs were used to discover the best feature set. From an initial set of 352 features, their GA determined that the centroid, fundamental frequency, energy, standard deviation and skewness of spectrum, and the amplitudes of the first two harmonics were the best features to achieve a successful classification rate. In a more recent work (Fujinaga & MacMillan, 2000) two additional significant features were reported: spectral irregularity and a modified version of tristimulus. Unfortunately, the selection of best features was heavily instrument-dependent. This problematic dependence has been also noted by other studies.

The intensive study of feature selection performed by Kostek (1998) represents another interesting approach. Kostek thoroughly examined approximately a dozen features. Examined features include, for example, energy of fundamental and of sets of partials, brightness, odd/even partials

ratio, tristimulus, and time delays of partials with respect to the fundamental. Kostek also explored, in other studies, the use of features derived from Wavelet Transforms instead of FFT-derived features. She found that the latter provided slightly better results than the former.

One of the more interesting aspects of Kostek’s work is her use of *rough sets* (Pawlak, 1982; Pawlak, 1991). Rough sets are a technique that was developed in the realm of knowledge-based discovery systems and data mining. Rough sets are implemented with the aim of classifying objects and then evaluating the relevance of the features used in the classification process. An elementary introduction to rough sets can be found in (Pawlak, 1998). We will return later with a fuller explication of rough sets.

Applications of the rough sets technique to different problems, including those of signal processing, can be found in (Czyzewski, 1998). Polkowski and Skowron (1998) present a thoughtful discussion of software tools implementing this kind of formalisms. Several studies by Kostek and her collaborators (Kostek, 1995; Kostek, 1998; Kostek, 1999; Kostek & Czyzewski, 2001) and by Wieczorkowska (Wieczorkowska, 1999b) used rough sets for reducing a large initial set of features for instrument classification. Wieczorkowska’s study provides the clearest example of set reduction using rough sets. She found that a starting set of sixty-two spectral and temporal features describing attack, steady state, and release of sounds could be further reduced to a set of sixteen features. Examples of the more significant features include: tristimulus, energy of 5<sup>th</sup>, 6<sup>th</sup> and 7<sup>th</sup> harmonics, energy of even partials, energy of odd partials, the most deviating of the lower partials, mean frequency deviation for low partials, brightness, and energy of high partials.

Temporal differences between values of the same feature have been rarely used in the reviewed studies. *Soundfisher*, the commercial system mentioned earlier, incorporates temporal differences alongside such basic features as loudness, pitch, brightness, bandwidth, and MFCCs (Wold, Blum, Keislar, & Wheaton, 1999). The Fujinaga or Eronen studies (cited above) have also incorporated temporal differences. To summarize this section, there are two inter-related factors that influence the success of feature-based identification and classification tasks. First, one must determine, and then select, the most discriminatory features from a seemingly infinite number of candidates. Second, one must reduce the number of applied features in order to make the resultant calculations tractable. We might intuitively conclude that using more than fifteen or twenty features seems to be a non-optimal strategy for attempting automatic classification of musical instruments. In order to settle into a short feature list, reliable data reduction techniques should be used. PCA and some types of Discriminant Analysis (both explained below) are robust and relatively easy to compute. Other techniques such as Kohonen maps, Genetic Algorithms, Rough Sets, etc., might yield better results when appropriate parameters and data are selected, but are inherently more complex. It is also clear that there are some features that are discriminative

only for certain types of instruments, and that not only temporal and spectral features, but also their temporal evolution, should be considered.

## 4. Techniques for sound classification

### 4.1 Perceptual-based clustering and classification

Retrieving sounds from a database by directly selecting signal features as those cited in the previous section is not a friendly task. As a consequence, exploiting relationships between them and high-level descriptions such as class or property (roughness, brightness) is required. A different way of retrieving sounds is by providing examples that are similar to what we are searching for; this is known as “query by example.” A specific kind of “query by example” is the one based on similarity of perception of sounds, instead of being based on sound categories. Leaving pitch, loudness and duration apart, this points directly to the notion of timbre and therefore to “timbre similarity.”

Several authors have proposed a measure of timbre similarity that has been derived from psycho-acoustical experiments (see section 2). This measure allows one to approximate the average judgment of perceived similarity obtained from people’s dissimilarity judgments between pairs of sounds. In order to do that, features or combinations of them, are used, with a possible weighting, to position the sound into a multi-dimensional space. Giving two sounds, a measure of timbre similarity can be approximated. Therefore, for a given target sound, it is possible to find in a database the one that “sounds” the closest to the target.

Misdariis et al. (1998) derived such a similarity measure approximation from Krumhansl (1989) and McAdams et al. (1995) experiments. Its formulation uses four features: log-attack-time, spectral centroid, spectral irregularity and spectral flux. Use of the similarity measure proposed by Misdariis et al. (1998) can be found, for example, in the search engine of IRCAM’s “Studio On Line” sound database. Peeters et al. (2000) proposed a new approximation adding the new feature “spectral spread.” They also proposed an equivalent approximation for percussive sounds derived from the Lakatos (2000) experiment. This latter uses the log-attack time, the spectral centroid and the temporal centroid.

A still remaining problem concerns the applicability of such a timbre similarity measure for sounds belonging to different families (as for example comparing a sustained harmonic sound – i.e. an oboe sound- with a percussive sound – i.e., a snare sound-). Current research is trying to construct a meta-timbre-space allowing such comparison between sounds belonging to different sound classes. Another kind of approach is that of Feiten and Günzel (1994), Così, De Poli, and Lauzzana (1994), or Spevak and Polfreman (2000). Signal features used in these works try to take into account the properties of human perception: MFCCs, Loudness critical-band rate time patterns, Lyon’s cochlear model, Gammatone filter banks, etc. These features are then used in

order to construct, automatically, what is called a “physical timbre space.” The “physical timbre space” aims at being the equivalent to usual timbre spaces but derived from signal features instead of from dissimilarity judgments yielded by human subjects in experimental conditions.

A “physical timbre space” can be derived from signal features using various techniques: Hierarchical Clustering, Multi-Dimensional Scaling analysis (see section 2) Kohonen Feature Maps (a.k.a. Self Organizing Maps, see note 8) or Principal Component Analysis (see section 3.3). Prandoni (1994) and De Poli and Prandoni (1997) used a combination of MFCCs, Self-Organized Maps, and PCA analysis. The authors applied this framework to the sounds of Wessel et al. (1987) and found that brightness and spectral slope are the features that best explain two of its “physical timbre space” axes. Prandoni (1994) used the barycentre of the representation of each sound family in a feature (MFCCs) space. Using MDS and Hierarchical Clustering analysis he found similar results than Grey did, and assigned the first two axes of his space to brightness and to something called “presence,” which is a measure of the energy inside the 800 Hz. region. In these two studies the obtained spaces were compared to usual timbre spaces coming from human experiments such as the above cited (sections 2 and 3.2). In Feiten and Günzel (1994) and Spevak and Polfreman (2000) the obtained spaces are used to make a temporal model of the sound evolution. The former authors define two sound feature maps (SFM). The first SFM is derived directly from a Kohonen Feature Map training using the MFCCs. This SFM, called the Steady State SFM, represents the steady parts of the sounds. Each sound is then represented by a trajectory between the states of the Steady State SFM. A Dynamic State SFM is then computed from these trajectories. The latter authors, on the other hand, make a comparison between different feature sets (Lyon’s cochlear model, Gammatone filterbank and MFCCs) considering their abilities to represent clear and separated trajectories in the SFM. They conclude that the best feature set is the Gammatone filterbank combined with Meddis’s inner hair cell model.

### 4.2 Taxonomic classification

In this section we are going to present different techniques that have been used for learning to classify isolated musical notes into instrument or music family categories. Although we have focused on the testing phase success rate as a way for evaluating them, we have to be cautious because other factors (number of instances used in the learning phase, number of instances used in the testing phase, testing procedure, number of classes to be learned, etc.) may have a large impact on the results.

#### *K-Nearest Neighbours*

The *K-Nearest Neighbours* (K-NN) algorithm is one of the most popular algorithms for instance-based learning. It first

stores the feature vectors of all the training examples and then, for classifying a new instance, it finds a set of  $k$  nearest training examples in the feature space, and assigns the new example to the class that has more examples in the set. Traditionally, the Euclidean distance measure is used to determine similarity. Although it is an easy algorithm to implement, the K-NN technique has several significant drawbacks:

- As it is a lazy algorithm (Mitchell, 1997), it requires having all the training instances in memory in order to yield a decision for classifying a new instance.
- It does not provide a generalization mechanism (because it is only based on local information).
- It is highly sensitive to irrelevant features that can dominate the distance metrics.
- It may require a significant computational load each time a new query is processed.

A k-NN algorithm classified 4 instruments with almost complete accuracy in Kaminskyj and Materka (1995) but the small size of the database (with restricted note range to one octave, although including different dynamics) was a drawback for taking this result as robust. In recent years Kaminskyj (2001) has reported hit rates of 82% for a database of 517 sounds and 19 instrumental categories. Some interesting features of this study are the use of PCA for reduction of data obtained after applying a Constant Q Transform and the use of a “reliability” estimation that can be extracted from confusion matrices. Martin and Kim (1998) developed a classification system that used a k-NN on a database of 1023 sounds with 31 features extracted from cochleograms (see also Martin, 1999). Their study included a hierarchical procedure consisting of:

- An initial discrimination of *pizzicati* from continuous notes.
- A discrimination between different “families” (e.g., sustained sounds further divided into strings, woodwind, and brass)
- A final classification of sounds into instrument categories.

When no hierarchy was used, Martin and Kim achieved a 87% classification success rate at the family level and a 61% rate at the instrument level. Use of the hierarchical procedure increased the accuracy at the instrument level to 79% but it degraded the performance at the family level to 79%. In the case of not including the hierarchical procedure, performance figures were lower than the ones they obtained with a Bayesian classifier. Similar results (65% for 27 instrument classes; 77% for a two-level 6-element hierarchy) were reported by Agostini et al. (2001). In this report, the k-NN technique compared unfavorably against Discriminant Functions and also against Support Vector Machines.

Eronen and Klapuri (2000) used a combination of k-NN and a Gaussian classifier (which was only used for rough discrimination between *pizzicati* and sustained sounds) for classifying 1498 samples into specific instrumental families or

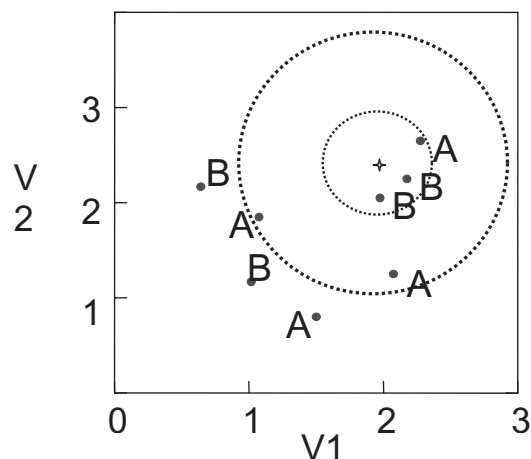


Fig. 3. An illustration of the K-NN technique. The point marked with a star would be classified as belonging to category “B” when  $K = 3$  (as two out of its 3 neighbours are from class “B”; but note that in case of using  $K = 5$  the point would be classified as “A” because there are 3 nearest neighbours belonging to this category and only 2 belonging to “B.”

specific instrument labels. Using a system architecture very similar to Martin and Kim’s hierarchy – wherein sounds are first classified in broad categories and then the classification is refined inside that category – they reported success rates of 75% in individual instrument classification and 94% for family classification. They also reported a small accuracy improvement by only using the best features for each instrument and no hierarchy at all (80%). A quite surprising result is the extreme degradation of performance results (35%) that has been reported in a more recent paper (Eronen, 2001). The explanation may be found in several facts: they used a larger and more varied database (5286 sounds coming from different collections) and more restrictive cross-validation methods (the test phase used sounds that were completely excluded from the learning set).

A possible enhancement of the K-NN technique, which includes the weighting of each feature according to its particular relevance for the task, has been used by the Fujinaga team (Fujinaga et al., 1998; Fujinaga, 1998; Fraser & Fujinaga, 1999; Fujinaga & MacMillan, 2000). In a series of three experiments using over 1200 notes from 39 different instruments, the initial success rate of 50%, observed when only the spectral shape of steady-state notes was used, increased to 68% when tristimulus, attack position, and features of the dynamically changing spectrum envelope (i.e., the change rate of the centroid) were added. In their last paper, a real-time version of this system was reported.

The k-NN literature – including the works of such research leaders as Martin and Fujinaga – consistently reports accuracy rates around 80%. Provided that the feature

selection has been optimized with genetic or other optimization techniques, one can thus interpret the 80% accuracy value as an estimation of the limitations of the K-NN algorithm. Therefore, more powerful techniques should be explored.

#### *Naïve Bayesian Classifiers*

A *Naïve Bayesian Classifier* (NBC) incorporates a learning step in which the probabilities for the classes and the conditional probabilities for a given feature and a given class are estimated. Probability estimates for each of these are based on their frequencies as found in a collection of training data. The set of these estimates corresponds to the learned hypothesis, which is formed by simply counting the occurrences of various data combinations within the training examples. Each new instance is classified based upon the conditional probabilities calculated during the learning phase. This type of classifier is called *naïve* because it assumes the independence of the features.

Brown (1999) used the NBC technique in conjunction with 18 Cepstral Coefficients computed after a constant Q transform. After clustering the feature vectors with a K-means algorithm, a Gaussian mixture model from their means and variances was built. This model was used to estimate the probabilities for a Bayesian classifier. It then classified 30 short sounds of oboe and sax with an accuracy rate of 85%. In a more recent paper (Brown, Houix, & McAdams, 2001) she and her collaborators reported similar hit rates for four classes of instruments (oboe, sax, clarinet and flute); these good results were replicated for different types of descriptors (cepstral coefficients, bin-to-bin differences of the constant-Q spectrum, and autocorrelation coefficients).

Martin (1999) enhanced a similar Bayesian classifier with context-dependent feature selection procedures, rule-one-out category decisions, beam search, and Fisher discriminant analysis, to estimate the maximum *a priori* probabilities. In (Martin & Kim, 1998) performance of this system was better than that of a K-NN algorithm at the instrument level with a 71% accuracy rate and equivalent to it at the family level with 85% accuracy rate.

Kashino and his team (1995) have also used a Bayesian classifier in their CASA system. Their implementation is reported to be able to classify, and even separate, five different instruments: clarinet, flute, piano, trumpet and violin. Unfortunately, no specific performance data are provided in their paper.

#### *Discriminant Analysis*

Classification using categories or labels that have been previously defined can be done with the help of *Discriminant Analysis* (DA) a technique that is related to multivariate analysis of variance (MANOVA) and multiple regression. DA attempts to minimize the ratio of within-class scatter to the between-class scatter and builds a definite decision region

between the classes. It provides linear, quadratic or logistic functions of the variables that “best” separate cases into two or more predefined groups. DA is also useful for determining which are the most discriminative features and the most similar/dissimilar groups. Surprisingly there have been very few studies using these techniques. Martin and Kim (1998) made limited use of this method when they used a linear DA to estimate the mean and variance of the Gaussian distributions of each class to be fed into an enhanced naive Bayesian classifier.

More recently Agostini et al. (2001) have found that a set of quadratic discriminant functions outperformed even Support Vector Machines (93% versus 70% hit rates) in classifying 1007 tones from 27 musical instruments with a very small set of descriptors. In our laboratory we carried out, some time ago, an unpublished study with 120 sounds from 8 classes and 3 families in which we got a 75% accuracy using also quadratic linear discriminant functions in two steps (sounds were first assigned to a family, and then they were specifically classified). As the features we used were not optimized for instrument classification but for perceptual similarity classification, it would be reasonable to expect still better results when including other more task-specific features. In a more recent work (Herrera et al., 2002) that used a database of 464 drum sounds (kick, snare, hihat, tom, cymbals) and an initial set of more than thirty different features, we got hit rates higher than 94% with four canonical Discriminant functions<sup>9</sup> that combined 18 features comprising some MFCCs, attack and decay descriptors, and relative energies in some selected bands.

#### *Higher Order Statistics*

When signals have Gaussian density distributions, we can describe them thoroughly with such second order measures as the autocorrelation function or the spectrum. In the case of noisy signals such as engine noises of sound effects, the variations in the spectral envelope do not allow a good signal characterisation and matching. A method to match signals using a variant of matched filter using polyspectral matching was presented in (Dubnov & Tishby, 1997), and it could be specifically useful for the classification of sounds from percussive instruments. There are some authors who claim that musical signals, because they have been generated through non-linear processes, do not fit a Gaussian distribution. In that case, using *higher order statistics* or polyspectra, as for example skewness of bispectrum and kurtosis of trispectrum, it is possible to capture all information that could be lost if using a simpler Gaussian model. With these techniques, and using a Maximum Likelihood classifier, Dubnov, Tishby, and Cohen (1997) have showed that discrimination between 18

<sup>9</sup>A canonical Discriminant function uses standardized values and Mahalannobis distances instead of raw values and Euclidean distances.



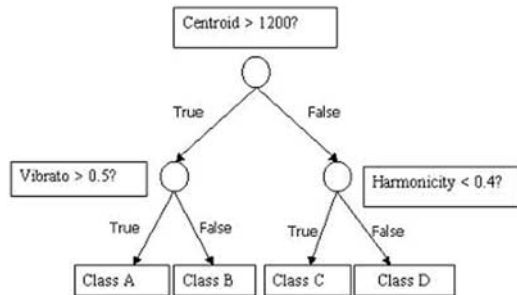


Fig. 4. An imaginary binary tree for classification of sounds into 4 different classes.

instruments from string, woodwind and brass families is possible. Unfortunately the detailed data that is presented there come from a classification experiment that used machine and other types of non-instrumental sounds. Acoustic justification for differences in kurtosis among families of instruments was provided in Dubnov and Rodet (1997). The measure of kurtosis was shown to correspond to the phenomenon of phase coupling, which implies coherence in phase fluctuations among the partials.

#### Binary trees

*Binary Trees*, in different formulations, are pervasively used for different machine learning and classification tasks. They are constructed top-down, beginning with the feature that seems to be the most informative one, that is, the one that maximally reduces entropy. Branches are then created from each one of the different values of this descriptor. In the case of non-binary valued descriptors, a procedure for dichotomic partitioning of the value range must be defined. The training examples are sorted to the appropriate descendant node, and the entire process is then repeated recursively using the examples of one of the descendant nodes, then with the other. Once the tree has been built, it can be pruned to avoid overfitting and to remove secondary features. Although building a binary tree is a recursive procedure, it is order of times faster than, for example, training a neural network.

Binary trees are best suited for approximating discrete-valued target functions but they can be adapted to real-valued features. Jensen and Arnsparang's binary decision tree (1999) exemplifies this approach to instrument classification. In their system, the trees are constructed by asking a large number of questions designed in each case to divide the data into two sets (e.g., "Is attack time longer than 60ms?"). Goodness of split (e.g., average entropy) is calculated and the question that renders the best goodness is chosen. Once the tree has been built using the learning set, it can be used for classifying new sounds because each leaf corresponds to one specific class. The tree can also be used for making explicit rules about which features better discriminate one instrument

from another. Unfortunately, detailed results regarding the classification of new sounds have not yet been published. Consult Jensen's thesis (1999), however, for his discussion of log-likelihood classification functions.

Wiczorkowska (1999a) used a binary tree approach, called the *C4.5* algorithm (Quinlan, 1993), to classify a database of 18 classes and 62 features. Accuracy rates varied between 64% and 68% depending on the test procedure applied. In our above-mentioned drum sounds classification study (Herrera et al., 2002) we obtained slightly better figures (83% of hit rates) using the *C4.5* algorithm for classifying nine different classes of instruments.

A final example of a binary tree for audio classification, although not specifically tested with musical sounds, is that of Foote (1997). His tree-based approach uses MFCCs and supervised vector quantization to partition the feature space into a number of discrete regions. Each split decision in the tree involves comparing one element of the vector with a fixed threshold that is chosen to maximize the mutual information between the data and the associated human-applied class labels. Once the tree is built, it can be used as a classifier by computing histograms of frequencies of classes in each leaf of the tree; histograms are similarly generated for the test sounds then compared with tree-derived histograms.

#### Artificial Neural Networks

An *Artificial Neural Network* (ANN) is an information processing structure that is composed of a large number of highly interconnected processing elements – called neurons or units – working in unison to solve specific problems. Neurons are grouped into layers (usually called *input*, *output*, and *hidden*) that can be interconnected through different connectivity patterns. An ANN learns complex mappings between *input* and *output* vectors by changing the weights that interconnect neurons. These changes may proceed either *supervised* or *unsupervised*. In the supervised case, a teaching instance is presented to the ANN, it is asked to generate an output, this out is then compared with an expected "correct" output, and the weights are consequently changed in order to minimize future errors. In the unsupervised case, the weights "settle" into a pattern that represents the collection of input stimulus.

A very simple feedforward network with a backpropagation training algorithm was used in Kaminskyj and Materka (1995). The network (a system with 3 input units, 5 hidden units, and 4 output units) learned to classify sounds from 4 very different instruments – piano, marimba, accordion and guitar – with an accuracy rate as high as 97%. Slightly better results were obtained, however, using a simpler K-NN algorithm. A three-way evaluative investigation involving a multilayer network, a time-delayed network, and a hybrid self-organizing network/radial basis function (see note 8) can be found in Cemgil and Gürgen (1997). Although very high success rates were found (e.g., 97% for the multilayer network, 100% for the time-delay network, and 94% for the

self-organizing network) it should be noted that the experiments used only 40 sounds from 10 different classes with the pitch range limited to one octave.

Implementations of self-organizing maps (Kohonen, 1995) can be found in Feiten and Günzel (1994); Cosi, De Poli, and Lauzzana (1994); Cosi et al. (1994); Toivainen et al. (1998). All these studies used some kind of human auditory preprocessing simulation to derive the features that were fed to the network. Each then built a map and evaluated its quality by comparing the network clustering results to those human-based sound similarity judgments (Grey, 1977; Wessel, 1979). From their maps and their comparisons they advance timbral spaces to be explored, or confirm/reject theoretical models that explain the data. We must note, however, that the classification we get from self-organizing maps has not traditionally been directly usable for instrument recognition, as the maps are not provided with any *a priori* label to be learned (i.e., no instrument names). Nevertheless, there are several promising mechanisms being explored for associating the output clusters to specific labels (e.g., the radial basis function used by Cemgil, (see above). The ARTMAP architecture (Carpenter, Grossberg, & Reynolds, 1991) is another means to implement this strategy. ARTMAP has a very complex topology including a couple of associative memory subsystems and also an “attentional” subsystem. Fragoulis et al. (1999) successfully used an ARTMAP for the classification of 5 instruments with the help of only ten features: slopes of the first five partials, time delays of the first 4 partials relative to the fundamental, and high frequency energy. The small 2% error rate reported was attributed to neglecting different playing dynamics in the training phase.

Kostek’s (1999) is the most exhaustive study on instrument classification using neural networks. Kostek’s team has carried out several studies (Kostek & Krolkowski, 1997; Kostek & Czyzewski, 2000; Kostek & Czyzewski, 2001) on network architecture, training procedures, and number and type of features, although the number of classes to be classified has been always too small. They have used a feedforward NN with one hidden layer. Initially their classes were instruments with somewhat similar sounds: trombone, bass trombone, English horn and contrabassoon. Lastly, papers with more categories (double bass, cello, viola, violin, trumpet, flute, clarinet . . .) have been added to the tests. Accuracy rates higher than 90% were achieved for different sets of four classes, although the results varied depending on the types of training and descriptors used.

Some ANN architectures are capable of approximating any function. This attribute makes neural networks a good choice when the function to be learned is not known in advance, or it is suspected to be non-linear. ANN’s do have some important drawbacks, however, that must be considered before they are implemented: the computation time for the learning phase is very long, adjustment of parameters can be tedious and prohibitively time consuming, and data overfitting can degrade their generalization capabilities. It is still an open question whether ANN’s can outperform simpler

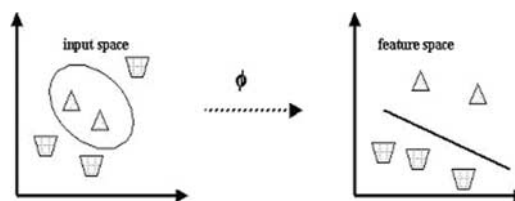


Fig. 5. In SVM’s the Kernel function  $\phi$  maps the input space (where discrimination of the two classes of instances is not easy to be defined) into a so-called feature space, where a linear boundary can be set between the two classes.

classification approaches. They do, however, exhibit one strong attribute that recommends their use: once the learning phase is completed, the classification decision is very fast when compared to other popular methods such as k-NN.

#### Support Vector Machines

SVMs are based on statistical learning theory (Vapnik, 1998). The basic training principle underlying SVMs is finding the optimal linear hyperplane such that the expected classification error for unseen test samples is minimized (i.e., they look for good generalization performance). According to the structural risk minimization inductive principle, a function that classifies the training data accurately, and which belongs to a set of functions with the lowest complexity, will generalize best regardless of the dimensionality of the input space. Based on this principle, a SVM uses a systematic approach to find a linear function with the lowest complexity. For linearly non-separable data, SVMs can (non-linearly) map the input to a high dimensional feature space where a linear hyperplane can be found. This mapping is done by means of a so-called *kernel* function (denoted by  $\phi$  in Fig. 5).

Although there is no guarantee that a linear solution will always exist in the high dimensional space, in practice it is quite feasible to construct a working solution. In other words, it can be said that training a SVM is equivalent to solving a quadratic programming with linear constraints and as many variables as data points. Anyway, SVM present also some drawbacks: first, there is a risk of selecting a non-optimal kernel function; second, when there are more than two categories to classify, the usual way to proceed is to perform a concatenation of two-class learning procedures; and third, the procedure is computationally intensive.

Marques (1999) used an SVM for the classification of 8 solo instruments playing musical scores from well-known composers. The best accuracy rate was 70% using 16 MFCCs and 0.2 second sound segments. When she attempted classification on longer segments an improvement was observed (83%). There were, however, two instruments found to be very difficult to classify: trombone and harpsichord. Another noteworthy feature of this study was the use of truly inde-

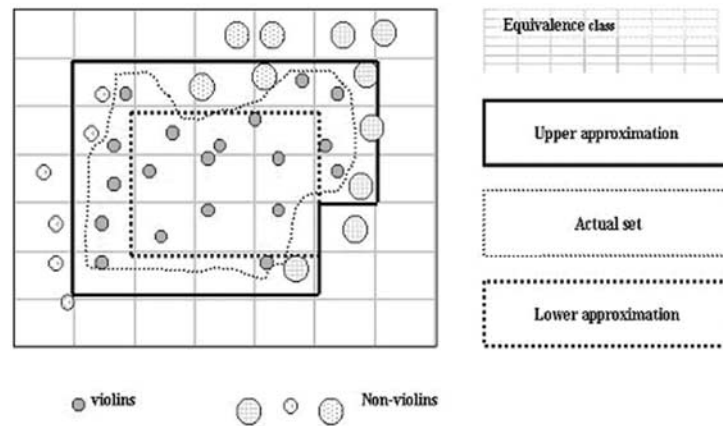


Fig. 6. An illustration of rough sets concepts.

pendent sets for the learning and for the testing consisting mainly of “solo” phrases from commercial recordings.

Agostini et al. have reported quite surprising results (Agostini et al., 2001). In their study an SVM performed marginally better than (Linear) Canonical Discriminant functions and also better than k-NN’s, but not nearly as good as a set of Quadratic Discriminant Functions (see section 4.2.3). Some promising applications of SVM that are related to music classification but are not specific to music instrument labelling can be found in Li and Guo (2000); Whitman, Flake; and Lawrence (2001); Moreno and Rifkin (2000); or Guo, Zhang, and Li (2001).

#### Rough Sets

*Rough sets* are a novel technique for evaluating the relevance of the features used for description and classification. These are similar to, but should not be confused with, *fuzzy sets*. In rough set theory, any set of similar or *indiscernible* objects is called an elementary set and forms a basic granule of knowledge about the universe; on the other hand, the set of *discernible* objects are considered rough (i.e., imprecise or vague). Vague concepts cannot be characterized in terms of information about their elements; however, they may be replaced by two precise concepts, respectively called the *lower approximation* and the *upper approximation* of the vague concept (see Fig. 6 for a graphical illustration of these ideas). The lower approximation consists of all objects that surely belong to the concept whereas the upper approximation contains all objects that could possibly belong to the concept. The difference between both approximations is called the *boundary region* of the concept.

The assignment of an object to a set is made through a membership function that has a probabilistic flavour. Once data are conveniently organized into information tables, this technique is used to assess the degree of vagueness of the

concepts and the interdependency of attributes; it therefore is useful for reducing complexity in the table without reducing the information it provides. Information tables regarding cases and features can be interpreted as conditional decision rules of the form **IF {feature x} is observed, THEN {is\_a\_Y\_object}**, and consequently they can be used as classifiers. When applied to instrument classification, Kostek (1998) reports accuracy rates higher than 80% for classification of the same 4 instruments mentioned in the ANN’s section. While both useful and powerful, the use of rough sets does entail some significant costs. The need for feature value quantization is the principal and non-trivial cost associated with rough sets. Furthermore, the choice of quantization method can affect output results. In the context of instrument classification, different quantization methods have been discussed in Kostek and Wiczorkowska (1997), Kostek (1998), and Wiczorkowska (1999b). When compared to neural networks or fuzzy sets rules, rough sets are computationally less expensive while at the same time yielding results similar to those obtained with the other two techniques.

#### Hidden Markov Models

Hidden Markov Models (HMMs) as the name implies, contain two components: a set of hidden variables that can not be observed directly from the data, and a Markov property that is usually related to some dynamical behaviour of the hidden variables.

A HMM is a generative model that assumes that a sequence of measurements or observations is produced through another sequence of hidden states  $s_1, \dots, s_T$ , so that the model generates, in each state, a random measurement drawn from a different (finite or continuous) distribution. Thus, given a sequence of measurements and assuming a certain sequence of hidden states, the HMM model specifies a joint probability distribution.

$$p(s_1 \dots s_T, x_1 \dots x_T) = p(s_1) p(x_1 | s_1) \prod_{t=2}^T p(s_t | s_{t-1}) p(x_t | s_t)$$

The HMM paradigm is used to solve three main tasks: classification, segmentation and learning. Learning is the first problem that needs to be solved in order to use a HMM model, unless the parameters of the model are externally specified. It means estimating the parameters of the models, usually iteratively done by the EM algorithm (Dempster, Laird, & Rubin, 1977). The tasks of segmentation and classification are accomplished via forward-backward recursions, which propagate information across the Markov state transition graph. The segmentation problem means finding the most likely sequence of the hidden states given an observation  $x_1 \dots x_T$ . Given several candidate HMM models that represent different acoustic sources (musical instruments in our case) the classification problem computes the probability that the observations came from these models. The model that gives the highest probability is chosen as the likely source of the observation.

HMMs have been used to address musical segmentation problems by several researchers (Raphael, 1999; Aucouturier & Sandler, 2001). These works dealt with segmentation of a sound into large-scale entities such as complete notes or sections of musical recordings, with the purpose of performing tasks such as score following or identification of texture changes in a musical piece.

Works that address the classification problem usually take a simpler view that discards the Markovian dynamics. Based on a work by Reynolds on speaker identification (Reynolds & Rose, 1995) several researchers considered a Gaussian Mixture Model (GMM) for computer identification of musical instruments (Brown, 1999; Marques, 1999). GMMs consider a continuous probability density of the observation, and model it as a weighted sum of several Gaussian densities. The hidden parameters in GMM are the mean vector, covariance matrix and mixture weight of the component densities. Parameter estimation is performed using an EM procedure or k-means. Using a GMM in an eight-instrument classification task, Marques reported an overall error rate of 5% for 32 Gaussians with MCCs as features. Brown performed a two-instrument classification experiment where she compared machine classification results with human perception for a set oboe and saxophone sounds. She reported a lower error rate for the computer than humans for oboe samples and roughly the same for the sax samples. Eronen and Klapuri (2000) also compare a GMM classifier to other classifiers for various features.

In the HMM model for sound clips presented by Zhang and Kuo (1998a; 1999b) they use a continuous observation density probability distribution function (pdf) with various architectures of the Markov transition graphs. They also incorporate an explicit State Duration model (semi-markov model, (Rabiner, 1989) for modelling the possibility that  $d$  consecutive observations belong to the same state. Denote a complete parameter set of HMM as  $\lambda = (A, B, D, \pi)$ , with  $A$  for the transition probability,  $B$  for the GMM parameters,  $D$

for duration pdf parameters and  $\pi$  for initial state distribution. In this model, two types of information are represented in the HMM: timbre and rhythm. Each kind of timbre is modelled by a state, and rhythm information is denoted by transition and duration parameters. The authors arrive at a three step learning procedure that first uses GMM for estimating  $B$ , then  $A$  is calculated from statistics of the state transitions and eventually  $D$  is estimated state by state, assuming a Gaussian density for the durations. This simplified procedure is not a strict HMM learning process and it is used to simplify the computational load of the learning stage. They report over 80% accurate classification rate for 50 sound clips, with misclassifications reportedly happening with classes of perceptually similar sounds, such as applause, rain, river and windstorm. The timbre of sound is described primarily by the frequency energy distribution that is extracted from short time spectrum. In their experiments, Zhang and Kuo employ a rather naive feature set for description of the timbre, that consist of log amplitude from a 128-point FFT vector (thus obtaining a 65 dimensional feature vector) calculated at approximately 9msec intervals. Depending on the type of sound that is analyzed, a partial or complete HMM model is employed. The simplest ones are single state sounds, and sounds that omit duration and transition information. These are used when every timbral state in the model can occur anywhere in time and for any duration. Second model includes transition probabilities, but without durations. The third (complete case) includes sounds such as footsteps and clock ticks, which carry both transition and duration information. An improvement to the timbral description was recently suggested by Casey and Westner (2001). Instead of using magnitude FFT, they suggest reduced rank spectra as a feature set for HMM classifier. After FFT analysis, singular value decomposition (SVD) is used to estimate a new basis for the data and, by discarding basis vectors with low eigenvalues, a data-reduction step is performed. Then the results are passed to independent component analysis (ICA<sup>10</sup>) which imposes additional constraints

<sup>10</sup>Independent component analysis (ICA) tries to improve upon the more traditional Principal Component Analysis (PCA) method of feature extraction by performing an additional linear transformation (rotating and scaling) of the PCA features so as to obtain maximal statistical independence between the feature vectors. One must note that PCA arrives at uncorrelated features, which are independent only when the signal statistics are Gaussian. It is claimed by several researchers that both in vision and sound the more "natural" features are the ICA vectors. The motivation for this claim is that ICA features are better localized in time (or space, in the case of vision) [Bell & Sejnowsky, 1996, 1997], and arrive at a more sparse representation of sound, that is, requiring less features, at every given instant of time (or space) in order to describe the signal. (One should note, though, that the total number of features needed to describe the whole signal is not changed). A serious study of the utility of ICA for sound recognition still needs to be carried out, especially in view of the computational overhead that needs to be "paid" for ICA processing, vs. the improvement in recognition rates.

on the output features. The resulting representation consists of a projection of a data into a lower-dimensional space with marginal distributions being approximately independent. They report a success rate of 92.65% for reduced-rank versus 60.61% for the full-rank spectra HMM classifier.

Another variant of Markov modelling, but this time using explicit (not hidden) observations with arbitrary length Markov modelling was used by Dubnov and Rodet (1998). In this work a universal classifier is constructed using a discrete set of features. The features were obtained by clustering (vector quantization of) cepstral and cepstral derivative coefficients. The motivation for this model is a universal sequence classification method of Ziv-Merhav (Ziv & Merhav, 1993) that performs matching of arbitrary sequences with no prior knowledge of the source statistics and having an asymptotic performance as good as any Markov or finite-state model. Two types of information are modelled in their work: timbre information and local sound dynamics, which are represented by cepstral and cepstral derivative features (observables). The long-term temporal behaviour is captured by modelling innovation statistics of the sequence, i.e., a probability to see a new symbol given the history of that sequence (for all possible length prefixes). By clever sampling of the sequence history, only most significant prefixes are used for prediction and clustering. The clustering method was tested on a set of 20 examples from 4 musical instruments, giving a 100% correct clustering.

## 5. Conclusions

We have examined the techniques that have been used for classification of isolated sounds and the features that have been found as more relevant for the task. We have also reviewed the perceptual features that account for clustering of sounds based on timbral similarity. Regarding the perceptual approach, we have presented empirical data for defining timbral spaces that are spanned by a small number of perceptual dimensions. These timbral spaces may help users of a music content-processing system to navigate through collections of sounds, to suggest perceptually based labels, and to perform groupings of sounds that capture similarity concepts. Regarding the taxonomic classification, we have discussed a variety of techniques and features that have provided different degrees of success when classifying isolated instrumental sounds. All of them show advantages and disadvantages that should be balanced according to the specifics of the classification task (database size, real-time constraints, learning phase complexity, etc.).

An approach yet to be tested is the combination of perceptual and taxonomic data in order to propose mixtures of perceptual and taxonomic labels (i.e., *bright snare-like tom* or *nasal violin-like flute*). It remains unclear, however, whether taxonomic classification techniques and features can be applied directly and successfully to the task of complex mixtures' *segmenting-by-instrument*. Additionally, because many of these techniques assume *a priori* isolation

of input sounds, they would not accomplish the requirements outlined by Martin (1999) for real-world sound-source recognition systems. Anyway, we have been lately focusing in a special type of sound mixtures, so-called "drum loops," where some dual and ternary combinations of sounds can be found, and we have obtained very good classification results adopting the isolated sounds approach (Herrera, Yeterian, & Gouyon, 2002). We have elsewhere (Herrera, Amatriain, Batlle, & Serra, 2000) suggested some strategies for overcoming this limitation and for guiding some forthcoming research.

## Acknowledgements

The writing of this paper was partially made possible thanks to funding received for the project CUIDADO from the European Community IST Program. The first author would like to express gratitude to Eloi Batlle and Xavier Amatriain for their collaboration as reviewers of preliminary drafts for some sections of this paper. He would also like to point out that large parts of this text have benefited from the editorial corrections and suggestions made by Stephen Downie, as editor of an alternative version, and by other three anonymous reviewers. The second author would also like to express gratitude to Stephen McAdams. Finally, thanks to Alex Sanjurjo for the graphical design of Figure 1.

## References

- Agostini, G., Longari, M., & Pollastri, E. (2001). Musical instrument timbres classification with spectral features. *IEEE Multimedia Signal Processing*, IEEE.
- American National Standards Institute. (1973). *American national psychoacoustical terminology*. S3.20. New York: American Standards Association.
- Antonic, D., & Zagar, M. (2000). Method for determining classification significant features from acoustic signature of mine-like buried objects. *15th World Conference on Non-Destructive Testing*. Rome, Italy.
- Aucouturier, J.J., & Sandler, M. (2001). Segmentation of musical signals using hidden markov models. *AES 110th Convention*.
- Bell, A.J., & Sejnowski, T.J. (1995). An information maximisation approach to blind separation and blind deconvolution. *Neural Computation*, 7, 1129–1159.
- Blum, A., & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97, 245–271.
- Brown, J.C. (1999). Musical instrument identification using pattern recognition with cepstral coefficients as features. *Journal of the Acoustical Society of America*, 105, 1933–1941.
- Brown, J.C., Houix, O., & McAdams, S. (2001). Feature dependence in the automatic identification of musical woodwind instruments. *Journal of the Acoustical Society of America*, 109, 1064–1072.

- Buller, G., & Lutman, M.E. (1998). Automatic classification of transiently evoked otoacoustic emissions using an artificial neural network. *British Journal of Audiology*, 32, 235–247.
- Carpenter, G.A., Grossberg, S., & Reynolds, J.H. (1991). ARTMAP: Supervised real-time learning and classification of nonstationary data by a self-organising neural network. *Neural Networks*, 4, 565–588.
- Casey, M.A., & Westner, A. (2001). Separation of mixed audio sources by independent subspace analysis. *Proceedings of the International Computer Music Conference, ICMA*.
- Cemgil, A.T., & Güngen, F. (1997). Classification of musical instrument sounds using neural networks. *Proceedings of SIU97*. Bodrum, Turkey.
- Cosi, P., De Poli, G., & Lauzzana, G. (1994). Auditory modeling and self-organizing neural networks for timbre classification. *Journal of New Music Research*, 21, 71–98.
- Cosi, P., De Poli, G., & Prandoni, P. (1994). Timbre characterization with Mel-cepstrum and neural nets. *Proceedings of the 1994 International Computer Music Conference*, (pp. 42–45). San Francisco, CA: International Computer Music Association.
- Czyzewski, A. (1998). Soft processing of audio signals. In: L. Polkowski & A. Skowron (Eds.), *Rough Sets in Knowledge Discovery: 2: Applications, Case Studies and Software Systems*. (pp. 147–165). Heidelberg: Physica Verlag.
- De Poli, G., & Prandoni, P. (1997). Sonological models for timbre characterization. *Journal of New Music Research*, 26, 170–197.
- Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, (34), 1–38.
- Dubnov, S., & Rodet, X. (1997). Statistical Modelling of Sound Aperiodicities. *Proceedings of International Computer Music Conference*, International Computer Music Association.
- Dubnov, S., & Rodet, X. (1998). Timbre recognition with combined stationary and temporal features. *Proceedings of 1998 International Computer Music Conference*. San Francisco, CA: International Computer Music Association.
- Dubnov, S., & Tishby, N. (1997). Analysis of sound textures in musical and machine sounds by means of higher order statistical features. *Proceedings of the International Conference on Acoustics Speech and Signal Processing*.
- Dubnov, S., Tishby, N., & Cohen, D. (1997). Polyspectra as measures of sound texture and timbre. *Journal of New Music Research*, 26, 277–314.
- Ellis, D.P.W. (1996). *Prediction-driven computational auditory scene analysis*. Unpublished doctoral dissertation, Massachusetts Institute of Technology, Cambridge, MA.
- Eronen, A. (2001). Comparison of features for musical instrument recognition. *2001 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'01)* IEEE.
- Eronen, A., & Klapuri, A. (2000). Musical instrument recognition using cepstral coefficients and temporal features. *Proceedings of the ICASSP*. Istanbul, Turkey: IEEE.
- Feiten, B., & Günzel, S. (1994). Automatic indexing of a sound database using self-organizing neural nets. *Computer Music Journal*, 18, 53–65.
- Foote, J.T. (1997). A similarity measure for automatic audio classification. *Proceedings of the AAAI 1997 Spring Symposium on Intelligent Integration and Use of Text, Image, Video, and Audio Corpora*. Stanford, CA: AAAI Press.
- Fragoulis, D.K., Avaritsiotis, J.N., & Papaodysseus, C.N. (1999). Timbre recognition of single notes using an ARTMAP neural network. *Proceedings of the 6th IEEE International Conference on Electronics, Circuits and Systems*. Paphos, Cyprus.
- Fraser, A., & Fujinaga, I. (1999). Toward real-time recognition of acoustic musical instruments. *Proceedings of the 1999 International Computer Music Conference*, 175–177. San Francisco, CA: International Computer Music Association.
- Fristrup, K.M., & Watkins, W.A. (1995). Marine animal sound classification. *Journal of the Acoustical Society of America*, 97, 3369–3370.
- Fujinaga, I. (1998). Machine recognition of timbre using steady-state tone of acoustical musical instruments. *Proceedings of the 1998 International Computer Music Conference*, (pp. 207–210). San Francisco, CA: International Computer Music Association.
- Fujinaga, I., & MacMillan, K. (2000). Realtime recognition of orchestral instruments. *Proceedings of the 2000 International Computer Music Conference*, (pp. 141–143). San Francisco, CA: International Computer Music Association.
- Fujinaga, I., Moore, S., & Sullivan, D.S. (1998). Implementation of exemplar-based learning model for music cognition. *Proceedings of the International Conference on Music Perception and Cognition*, (pp. 171–179).
- Gorman, R.P., & Sejnowski, T.J. (1988). Learned classification of sonar targets using a massively parallel network. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 36, 1135–1140.
- Grey, J.M. (1977). Multidimensional perceptual scaling of musical timbres. *Journal of the Acoustical Society of America*, 61, 1270–1277.
- Grey, J.M. (1978). Timbre Discrimination in Musical Patterns. *Journal of the Acoustics Society of America*, 64, 467–472.
- Grey, J.M., & Gordon, J.W. (1978). Perceptual effects of spectral modifications on musical timbres. *Journal of the Acoustical Society of America*, 63, 1493–1500.
- Guo, G.D., Zhang, H.J., & Li, S.Z. (2001). Boosting for content-based audio classification and retrieval: An evaluation. *IEEE International Conference on Multimedia and Expo*.
- Herre, J., Allamanche, E., & Hellmuth, O. (2001). Robust matching of audio signals using spectral flatness features. *2001 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'01)* IEEE.
- Herrera, P., Amatriain, X., Batlle, E., & Serra, X. (2000). Towards instrument segmentation for music content description: A critical review of instrument classification techniques. *International Symposium on Music Information Retrieval*.

- Herrera, P., Amatriain, X., Batlle, E., & Serra, X. (2002). A critical review of automatic musical instrument classification. In: D. Byrd, J.S. Downie, & T. Crawford (Eds.), *Recent Research in Music Information Retrieval: Audio, MIDI, and Score*. Dordrecht, Netherlands: Kluwer Academic Press.
- Herrera, P., Yeterian, A., & Gouyon, F. (2002). *Automatic classification of drum sounds: a comparison of feature selection methods and classification techniques*. International Conference on Music and Artificial Intelligence. Edinburgh, United Kingdom.
- Jensen, K. (1999). *Timbre models of musical sounds*. Unpublished doctoral dissertation, University of Copenhagen, Copenhagen, Denmark.
- Jensen, K., & Arnspang, J. (1999). Binary decision tree classification of musical sounds. *Proceedings of the 1999 International Computer Music Conference*. San Francisco, CA: International Computer Music Association.
- Kaminskyj, I. (2001). Multi-feature Musical Instrument Sound Classifier. *Australasian Computer Music Conference*.
- Kaminskyj, I., & Materka, A. (1995). Automatic source identification of monophonic musical instrument sounds. *Proceedings of the IEEE International Conference On Neural Networks, 1*, 189–194.
- Kaminskyj, I., & Voumard, P. (1996). Enhanced automatic source identification of monophonic musical instrument sounds. *Proceedings of the 1996 Australian New Zealand Conference on Intelligent Information Systems*, (pp. 76–79).
- Kartomi, M. (1990). *On Concepts and Classification of Musical Instruments*. Chicago: The University of Chicago Press.
- Kashino, K., & Murase, H. (1997a). A music stream segregation system based on adaptive multi-agents. *Proceedings of the 15th International Joint Conference on Artificial Intelligence (IJCAI-97) 2*, 1126–1131.
- Kashino, K., & Murase, H. (1997b). Sound source identification for ensemble music based on the music stream extraction. *Working Notes of the IJCAI-97 Computational Auditory Scene Analysis Workshop*, 127–134.
- Kashino, K., Nakadai, K., Kinoshita, T., & Tanaka, H. (1995). Application of Bayesian probability network to music scene analysis. *Proceedings of the 1995 International Joint Conference on Artificial Intelligence*, (pp. 52–59). Montreal, Canada.
- Keislar, D., Blum, T., Wheaton, J., & Wold, E. (1995). Audio analysis for content-based retrieval. *Proceedings of the 1995 International Computer Music Conference*, (pp. 199–202). San Francisco, CA: International Computer Music Association.
- Keislar, D., Blum, T., Wheaton, J., & Wold, E. (1999). A contentware sound browser. *Proceedings of the 1999 International Computer Music Conference*. San Francisco, CA: International Computer Music Association.
- Kohonen, T. (1995). *Self-organizing maps*. Berlin: Springer-Verlag.
- Kostek, B. (1995). Feature extraction methods for the intelligent processing of musical sounds. *AES 100th convention*, Audio Engineering Society.
- Kostek, B. (1998). Soft computing-based recognition of musical sounds. In: L. Polkowski & A. Skowron (Eds.), *Rough Sets in Knowledge Discovery*. Heidelberg: Physica-Verlag.
- Kostek, B. (1999). *Soft computing in acoustics: Applications of neural networks, fuzzy logic and rough sets to musical acoustics*. Heidelberg: Physica Verlag.
- Kostek, B., & Czyzewski, A. (2000). An approach to the automatic classification of musical sounds. *AES 108th convention*. Paris: Audio Engineering Society.
- Kostek, B., & Czyzewski, A. (2001). Representing musical instrument sounds for their automatic classification. *Journal of the Audio Engineering Society*, 49, 768–785.
- Kostek, B., & Krolkowski, R. (1997). Application of artificial neural networks to the recognition of musical sounds. *Archives of Acoustics*, 22, 27–50.
- Kostek, B., & Wiczorkowska, A. (1997). Parametric representation of musical sounds. *Archives of Acoustics*, 22, 3–26.
- Krimphoff, J., McAdams, S., & Winsberg, S. (1994). Caractérisation du timbre des sons complexes. II: Analyses acoustiques et quantification psychophysique. *Journal de Physique*, 4, 625–628.
- Krumhansl, C.L. (1989). Why is musical timbre so hard to understand? In: S. Nielzenand & O. Olsson (Eds.), *Structure and perception of electroacoustic sound and music* (pp. 43–53). Amsterdam: Elsevier.
- Lakatos, S. (2000). A common perceptual space for harmonic and percussive timbres. *Perception and Psychophysics*, Submitted for publication.
- Li, S.Z., & Guo, G. (2000). Content-based audio Classification and retrieval using SVM learning. *Proceedings of the First IEEE Pacific-Rim Conference on Multimedia*. Sydney, Australia: IEEE.
- Liu, Z., Wang, Y., & Chen, T. (1998). Audio feature extraction and analysis for scene segmentation and classification. *Journal of VLSI Signal Processing*, 20, 61–79.
- Marques, J. (1999). *An automatic annotation system for audio data containing music*. Unpublished master's thesis, Massachusetts Institute of Technology, Cambridge, MA.
- Martin, K.D. (1999). *Sound-source recognition: A theory and computational model*. Unpublished doctoral dissertation, MIT, Cambridge, MA.
- Martin, K.D., & Kim, Y.E. (1998). Musical instrument identification: A pattern-recognition approach. *Proceedings of the 136th meeting of the Acoustical Society of America*.
- McAdams, S., & Winsberg, S. (in preparation). A meta-analysis of timbre space. I: Multidimensional scaling of group data with common dimensions, specificities, and latent subject classes.
- McAdams, S., Winsberg, S., de Soete, G., & Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject classes. *Psychological Research*, 58, 177–192.
- McAdams, S., Susini, P., Krimphoff, J., Peeters, G., Rioux, V., Misdariis, N., & Smith, B. (in preparation). A meta-analysis of timbre space. II: Psychophysical quantification of common dimensions.

- McLaughling, J., Owsley, L.M.D., & Atlas, L.E. (1997). Advances in real-time monitoring of acoustic emissions. *Proceedings of the SAE Aerospace Manufacturing Technology and Exposition*, (pp. 291–297). Seattle, Washington.
- Michie, D., Spiegelhalter, D.J., & Taylor, C.C. (1994). *Machine learning, neural and statistical classification*. Chichester: Ellis Horwood.
- Mills, H. Automatic detection and classification of nocturnal migrant bird calls. *Journal of the Acoustical Society of America*, *97*, 3370–3371.
- Misdariis, N., Smith, B., Pressnitzer, D., Susini, P., & McAdams, S. (1998). Validation and multidimensional distance model for perceptual dissimilarities among musical timbres. *Proc. of Joint meeting of the 16th congress on ICA, 135th meeting of ASA*.
- Mitchell, T.M. (1997). *Machine learning*. Boston, MA: McGraw-Hill.
- Moreno, P.J., & Rifkin, R. (2000). Using the Fisher Kernel method for web audio classification. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*.
- Pawlak, Z. (1982). Rough sets. *Journal of Computer and Information Science*, *11*, 341–356.
- Pawlak, Z. (1991). *Rough sets: Theoretical aspects of reasoning about data*. Dordrecht: Kluwer Academic Publishers.
- Pawlak, Z. (1998). Rough set elements. In: L. Polkowski & A. Skowron (Eds.), *Rough Sets in Knowledge Discovery*. Heidelberg: Physica-Verlag.
- Peeters, G., McAdams, S., & Herrera, P. (2000). Instrument sound description in the context of MPEG-7. *Proceedings of the 2000 International Computer Music Conference*. San Francisco, CA: International Computer Music Association.
- Pfeiffer, S.R., Lienhart, R., & Effelsberg, W. (1998). *Scene determination based on video and audio features* (TR-98-020). University of Mannheim, Mannheim, Germany.
- Plomp, R. (1970). *Old and new data on tone perception* (IZF1970-14).
- Plomp, R. (1976). *Aspects of Tone Sensation: A Psychophysical Study*. London: Academic Press.
- Polkowski, L., & Skowron, A. (1998). *Rough sets in knowledge discovery*. Heidelberg: Physica-Verlag.
- Potter, J.R., Mellinger, D.K., & Clark, C.W. (1994). Marine mammal call discrimination using artificial neural networks. *Journal of the Acoustical Society of America*, *96*, 1255–1262.
- Prandoni, P. (1994). *An analysis-based timbre space*.
- Quinlan, J.R. (1993). *C4.5: Programs for machine learning*. San Mateo, CA: Morgan Kaufmann.
- Rabiner, L.R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*, 257–286.
- Rabiner, L.R., & Juang, B.H. (1993). *Fundamentals of speech recognition*. New York: Prentice-Hall.
- Raphael, C. (1999). Automatic segmentation of acoustic musical signals using hidden Markov models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *21*, 360–370.
- Reynolds, D.A., & Rose, R.C. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, January 1995.
- Rochebois, T., & Charbonneau, G. (1997). Cross-synthesis using interverted principal harmonic sub-spaces. In: M. Leman (Ed.), *Music, Gestalt and Computing* (pp. 221–244). Berlin: Springer.
- Sandell, G.J., & Martens, W.L. (1995). Perceptual evaluation of principal-component-based synthesis of musical timbres. *Journal of the Acoustical Society of America*, *43*, 1013–1028.
- Scheirer, E.D. (2000). *Music-listening systems*. Unpublished doctoral dissertation, Massachusetts Institute of Technology.
- Schön, P.C., Puppe, B., & Manteuffel, G. (2001). Linear prediction coding analysis and self-organizing feature map as tools to classify stress calls of domestic pigs (*Sus scrofa*). *Journal of the Acoustical Society of America*, *110*, 1425–1431.
- Shiyong, Z., Zehan, C., Fei, G., Li, F., & Shouzong, X. (1998). The knowledge-based signal analysis for a heart sound information system. *Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, *20*, (pp. 1622–1624). Piscataway, NJ: IEEE Computer Society Press.
- Smoliar, S.W., & Wilcox, L.D. (1997). Indexing the content of multimedia documents. *Proceedings of the Second International Conference on Visual Information Systems*, (pp. 53–60). San Diego, CA.
- Spevak, C., & Polfremam, R. (2000). Analyzing auditory representations for sound classification with self-organizing neural networks. *COST G-6 Conference on Digital Audio Effects (DAFX-00)*.
- Toivainen, P., Tervaniemi, M., Louhivuori, J., Saher, M., Huottilainen, M., & Nääätänen, R. (1998). Timbre similarity: Convergence of neural, behavioral, and computational approaches. *Music Perception*, *16*, 223–241.
- Vapnik, V.N. (1998). *Statistical learning theory*. New York: Wiley.
- Varga, A.P., & Moore, R.K. (1990). Hidden Markov model decomposition of speech and noise. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 845–848.
- Wedin, L., & Goude, G. (1972). Dimension analysis of the perception of instrumental timbre. *Scandinavian Journal of Psychology*, *13*, 228–240.
- Wessel, D. (1979). Timbre space as a musical control structure. *Computer Music Journal*, *3*, 45–52.
- Wessel, D., Bristow, D., & Settel, Z. (1987). Control of phrasing and articulation in synthesis. *International Computer Music Conference*, (pp. 108–116). San Francisco: International Computer Music Association.
- Whitman, B., Flake, G., & Lawrence, S. (2001). Artist detection in music with Minnowmatch. *Proceedings of the 2001 IEEE Workshop on Neural Networks for Signal Processing*, 559–568.
- Wieczorkowska, A. (1999a). Classification of musical instrument sounds using decision trees. *Proceedings of the 8th*



- International Symposium on Sound Engineering and Mastering, ISSEM'99*, (pp. 225–230). Gdansk, Poland.
- Wieczorkowska, A. (1999b). Rough sets as a tool for audio signal classification. In: Z.W. Ras & A. Skowron (Eds.), *Foundations of Intelligent Systems: Proceedings of the 11th International Symposium on Foundations of Intelligent Systems (ISMIS-99)* (pp. 367–375). Berlin: Springer-Verlag.
- Wish, M., & Carroll, J.D. (1982). Multidimensional scaling and its applications. In: P.R. Krishnaiah & L.N. Kanal (Eds.), *Handbook of statistics: 2*. (pp. 317–345). Amsterdam: North-Holland.
- Wold, E., Blum, T., Keislar, D., & Wheaton, J. (1999). Classification, search and retrieval of audio. In: B. Furth (Ed.), *Handbook of Multimedia Computing* (pp. 207–226). Boca Raton, FL: CRC Press.
- Wold, E., Blum, T., Keislar, D., & Wheaton, J. (1996). Content-based classification, search and retrieval of audio. *IEEE Multimedia*, 3, 27–36.
- Zhang, T., & Jay Kuo, C.-C. (1998a). Content-based classification and retrieval of audio. *SPIE's 43rd Annual Meeting – Conference on Advanced Signal Processing Algorithms, Architectures, and Implementations VIII*, 3461, (pp. 432–443). San Diego, CA.
- Zhang, T., & Jay Kuo, C.-C. (1998b). Hierarchical system for content-based audio classification and retrieval. *SPIE's Conference on Multimedia Storage and Archiving Systems III*, 3527, (pp. 398–409). Boston: SPIE.
- Zhang, T., & Jay Kuo, C.-C. (1999a). Heuristic approach for generic audio data segmentation and annotation. *ACM Multimedia Conference*, (pp. 67–76). Orlando, FLA.
- Zhang, T., & Jay Kuo, C.-C. (1999b). Hierarchical classification of audio data for archiving and retrieving. *IEEE International Conference On Acoustics, Speech, and Signal Processing*, 6, 3004. Phoenix, AR.
- Ziv, J., & Merhav, N. (1993). A measure of relative entropy between individual sequences with application to universal classification. *IEEE Transactions on Information Theory*, (July) 1270–1279.

Gómez, E. & **Herrera, P.** (2008). "Comparative Analysis of Music Recordings from Western and Non-Western traditions by Automatic Tonal Feature Extraction". *Empirical Musicology Review*, 3(3), pp. 140-156.

DOI: <https://doi.org/10.18061/1811/34105>

Online ISSN: 1559-5749

# Comparative Analysis of Music Recordings from Western and Non-Western traditions by Automatic Tonal Feature Extraction

EMILIA GÓMEZ

*Music Technology Group, Universitat Pompeu Fabra  
Sonology Department, Escola Superior de Música de Catalunya*

PERFECTO HERRERA

*Music Technology Group, Universitat Pompeu Fabra  
Sonology Department, Escola Superior de Música de Catalunya*

**ABSTRACT:** The automatic analysis of large musical corpora by means of computational models overcomes some limitations of manual analysis, and the unavailability of scores for most existing music makes necessary to work with audio recordings. Until now, research on this area has focused on music from the Western tradition. Nevertheless, we might ask if the available methods are suitable when analyzing music from other cultures. We present an empirical approach to the comparative analysis of audio recordings, focusing on tonal features and data mining techniques. Tonal features are related to the pitch class distribution, pitch range and employed scale, gamut and tuning system. We provide our initial but promising results obtained when trying to automatically distinguish music from Western and non-Western traditions; we analyze which descriptors are most relevant and study their distribution over 1500 pieces from different traditions and styles. As a result, some feature distributions differ for Western and non-Western music, and the obtained classification accuracy is higher than 80% for different classification algorithms and an independent test set. These results show that automatic description of audio signals together with data mining techniques provide means to characterize huge music collections from different traditions and complement musicological manual analyses.

Submitted 2008 March 5; Accepted 2008 April 2; Reviewed 2008 August 20.

**KEYWORDS:** *audio description, tonality, machine learning, comparative analysis*

## INTRODUCTION

### Goals and motivation

THERE is a wealth of literature on music research that focuses on the comparative study of different musical genres, styles and traditions. According to Toiviainen and Eerola (2006, p. 1), “typical research questions in this area of inquiry involve the evolution of a musical style, typical musical features in the works a composer, or similarities and differences across music traditions from various geographical regions”. Traditional research methods have been mainly based on either manual analysis of notated scores or aural analysis of music recordings. Although these manual analyses provide very accurate and expert information, they might have two potential limitations, as pointed out in the work by Toiviainen and Eerola (2006). First, manual annotation is a time consuming task, and this makes these studies to be based on a relatively small music collection that might not be then representative, in a statistical sense, of the corpora in study. Second, manual annotations might be subjective or prone to errors, especially if they are generated by different people with slightly different criteria without a common methodology (Lesaffre et al., 2004).

One way to overcome these limitations is to introduce the use of computational methods, which allows automating (in different degrees) the analysis of large musical collections. Many recent studies have

been devoted to apply computational models to comparative music research. Toiviainen and Eerola (2006) provide a very good overview of computational approaches to comparative music research, including issues related to music representation, musical feature extraction and data mining techniques. They also provide some examples of visualization of large musical collections based on these techniques, focusing on the analysis of MIDI representations.

Some studies in the field of Music Information Retrieval (MIR) have been also devoted to apply these methods to the analysis of audio recordings, mainly in some applied contexts such as music genre classification or artist identification (e.g. Tzanetakis & Cook, 2002). Extracted features are related to different musical facets and a varied set of data mining techniques are afterward applied to this set of descriptors. Timbre and rhythmic features are the most commonly used ones. They provide a way to characterize differences in instrumentation and meter, which are usually enough to discriminate diverse musical styles. Timbre is usually characterized by a group of descriptors directly computed from the signal spectrum (Tzanetakis & Cook, 2002), and common rhythmic features are tempo and Inter-Onset-Intervals (IOIs) histograms, which represent the predominant pulses (Gouyon & Dixon, 2005).

These methods are also considered when trying to measure the similarity between two musical pieces based on different criteria (used instruments, rhythmic pattern or harmonic progression). The definition of a music similarity measure is a very complex and somehow subjective task.

Until now, MIR research has mainly focused on the analysis of music from the so-called “Western tradition”, given that most of MIR systems are targeted toward this kind of music. Nevertheless, we might ask if the available descriptors and techniques are suitable when analyzing music from different traditions. The term *Western* is generally employed to denote most of the cultures of European origin and most of their descendants, and it is often used in contrast to other cultures including Asians, Africans, Native Americans, Aborigines, Arabs, and prehistoric tribes. Tzanetakis et al. (2007) have recently introduced the concept of *Computational Ethnomusicology* to refer to the use of computer tools to assist in ethnomusicological research, providing some guidelines and specific example of this type of multidisciplinary research. In this context, we present here an example of the use of audio analysis tools for comparative analysis of music from different traditions and genres.

The goal of the present study is to provide an empirical approach to the comparative analysis of music audio recordings, focusing on tonal features and a music collection from different traditions and musical styles. These descriptors are related to the pitch class distribution of a piece, its pitch range or tessitura and the employed scale and tuning system, being the feature extraction process derived from mathematical models of Western musical scales and consonance. We provide our initial but promising results obtained when trying to automatically distinguish or classify music from Western and non-Western traditions by means of automatic audio feature extraction and data mining techniques. Having in mind this goal, we analyze which features might be relevant to this task and study their distribution over a large music collection. We then apply some data mining techniques intended to provide an automatic classification of a music recording into Western and non-Western categories. From an applied point of view, we investigate if it is possible to automatically classify music into Western and non-Western by just analyzing audio data.

### **Musical scales in Western and non-Western music**

As mentioned above, we hypothesize that tonal descriptors related to the pitch class distribution of a piece, pitch range and employed scale and gamut may be useful to differentiate music from different traditions and styles.

A *scale* is a small set of notes in ascending or descending order (usually within an octave), being a sequence long enough to define a mode, tonality or another linear construction which starts and ends on its main note. These scales establish the basis for melodic construction. On the other side, the *gamut* is defined as the full range of pitches in a musical system, appearing when all possible variants of all possible scales are combined.

Scales in traditional Western music generally consist of seven notes (i.e. scale degrees), repeat at the octave, and are separated by whole and half step intervals of tones and semitones. Western music in the Medieval and Renaissance periods (1100-1600) tends to use the diatonic scale C-D-E-F-G-A-B, with rare and unsystematic presence of accidentals. Music of the common practice period (1600-1900) uses three types of scales: the diatonic scale and the melodic and harmonic minor scales, having 7 notes. In the 19<sup>th</sup> and 20<sup>th</sup> centuries, additional types of scale are explored, as the chromatic (12 notes), the whole tone (6

notes), the pentatonic (5 notes), the octatonic or diminished scales (Drabkin, 2008). As a general observation, in traditional Western music, scale notes are most often separated by equally-tempered tones or semitones, creating a gamut of 12 pitches per octave, so that the frequency ratio between consecutive semitones is equal to  $st = \sqrt[12]{2}$ , i.e. the interval value in the logarithm ‘cent’ metric is equal to 100 cents.

Many other musical traditions employ scales that include other intervals or a different number of pitches. According to Burns (1998, p. 217), the use of discrete pitch relationships is largely universal, pitch glides (as glissandos or portamentos) are used as embellishment and ornamentation in most musical cultures, and the concept of octave equivalence, although far from universal in early and structurally simpler music, seems to be common to more advanced musical systems.

For instance, gamelan music uses a small variety of scales including Pélog and Sléndro, not including equally tempered intervals (Carterette & Kendall, 1994). Ragas in Indian classical music often employ intervals smaller than a semitone, as both musical systems of India (Hindustani and Karnatic) are based on 22 possible intervals per octave and are not equal interval. Arabic music may use quarter tone intervals. According to Burns (1998), there are different theories as to the number of used intervals (ranging from 15 to 24) and some controversy as to where they are true quarter tones or merely microtonal variations of certain intervals.

Central and southern African music is characterized by the dominance of rhythmic and percussive devices, and the scales of musical instruments do not seem to be an approximation of the Western tempered scale (Merriam, 1959; VV.AA, 1973). Chinese and Japanese tuning also differ from equal-tempered scale (Piggott, 1891-1892).

In addition to the mentioned musical traditions, there are also other musical genres (apart from classical Western music) that may employ scale intervals smaller than a semitone. For instance, the blue note is an interval that is neither major nor minor, but in between, giving it a characteristic flavor. In blues, a pentatonic scale is often used, and in jazz many different modes and scales are found (being chromatic scales commonly used), often in the same piece.

### MUSIC COLLECTION

A relevant step for a comparative study of music material is the definition of a proper audio collection. This collection should be representative of the different styles present in music from Western and non-Western tradition, which is an arduous task, given the variety of both categories.

We have tried to cope with the variety of both classes of music, gathering a music made of 500 audio recordings from non-Western music (distributed by region: Africa, Java, Arabic, Japan, China, India and Central Asia). These samples consist of recordings of traditional music from different areas, and we discarded those having some Western influence (equal-tempered instruments as the piano, for instance).

We also considered 1000 recordings from Western music, gathered from commercial CDs and distributed across the musical genres presented in Figure 1.

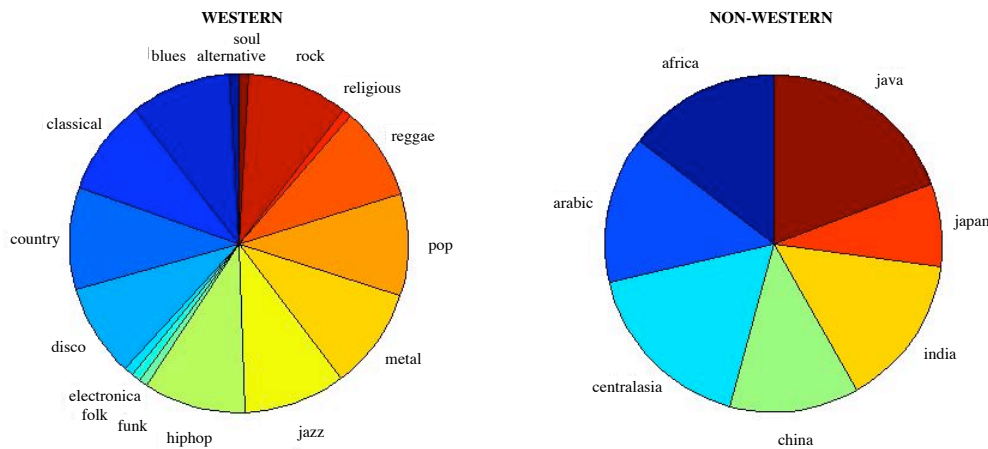


Fig. 1. Distribution of musical genres within the music collection

Non-Western music was chosen to be representative of the musical tradition of each geographical region, and Western music was chosen to cover a set of varied musical genres, which might be more representative of the different types of music than geographical ordering. The “Western” collection that was chosen has been widely used in automatic genre recognition (Tzanetakis, 2001; Holzapfel, 2007; Rentfrow, 2003).

## AUDIO FEATURE EXTRACTION

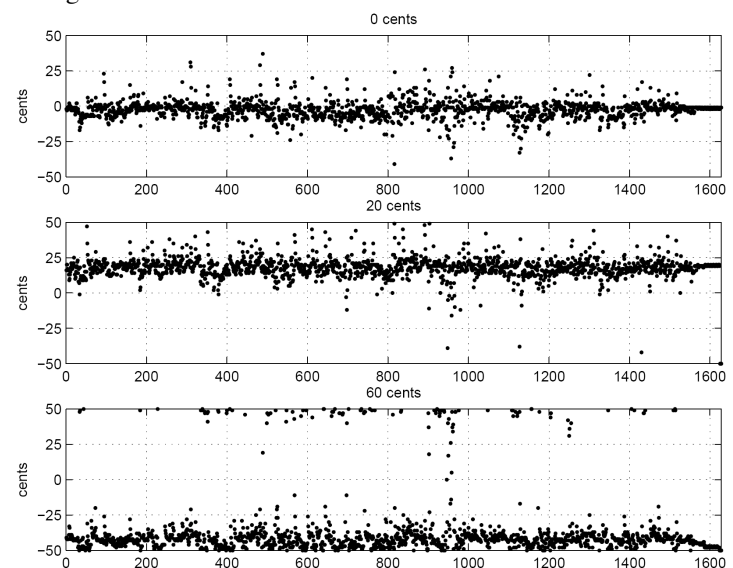
Another relevant task is the definition and computation of a representative set of musical features for the studied problem. These features should be automatically extractable from audio recordings.

Based on previous studies, we hypothesize that features derived from a tonal analysis of the piece might be relevant for comparative analysis and music similarity in this particular context, as they represent the pitch class distribution of the piece and they are not influenced by instrumentation or tempo (Toiviainen & Eerola, 2006; Gómez, 2006, p. 153-183).

We compute these features over the first 30 seconds of each musical piece, for two different reasons. First, 30 seconds are being considered enough in the MIR community to recognize a certain musical style and a musical key. According to Gómez (2006, p. 134), the accuracy rate of a key estimation algorithm is similar if we only consider the beginning of a piece (from 15 seconds) than the whole piece. Second, and from a practical point of view, the computation speed is reduced. We then assume that analyzing the first 30 seconds of a musical recording should be enough to classify if the piece belongs or not to the Western music tradition. In order to check this fact, we listened to the starting segments of all the pieces and discarded few non representative parts (containing silences or ambiguous introductions). We present here the set of audio features that has been considered in this study.

### Tuning frequency

One of the features that we consider relevant for this task is the frequency used to tune a musical piece, which is close to 440 Hz in the Western tradition. We estimate the tuning frequency (i.e. its difference with 440 Hz) as the value that minimizes the deviation of the main spectral peaks from an equal-tempered scale. These spectral peaks are obtained after frequency analysis of the audio signal through the Discrete Fourier Transform (DFT). An estimation of the tuning frequency is computed in a frame basis (frames have a 100 ms duration and are 50% overlapped), and a global estimate is derived from the frame values by means of a histogram. A more detailed description of the algorithm is presented in (Gómez, 2006, p. 71-76). An example is provided in Figure 2.



**Fig. 2.** Frequency deviation with respect to 440 Hz (in cents) vs. frame index, computed for a piece tuned to 440 Hz (up), with a deviation of 20 cents (middle) and 60 cents (down).

## High-resolution pitch class distributions

Pitch-class distributions from symbolic data are already considered in Toiviainen and Eerola (2006) for the comparative analysis of symbolic data from different geographical regions. We extend this idea to the analysis of audio recordings by computing pitch class distributions from audio signals.

We call the obtained features the Harmonic Pitch Class Profile (HPCP), and the procedure for its computation is presented in detail in Gómez (2006) and illustrated in Figure 3. The HPCP is computed in a frame basis, considering 100 ms overlapped frames. We perform a spectral analysis of each audio frame followed by a peak estimation procedure. The peak frequencies are then mapped into pitch-class values according to the tuning frequency value previously estimated. The HPCP vector is computed using an interval resolution of 10 cents per semitone, so that the vector size is equal to 120 points (10 values per semitone). This resolution is chosen in order to achieve a more detailed representation of the pitch class distribution of the piece and to cope with the small pitch variations obtained through different tuning systems and scales. Finally, we consider the HPCP average of the frames belonging to the considered audio excerpt (30 first seconds).

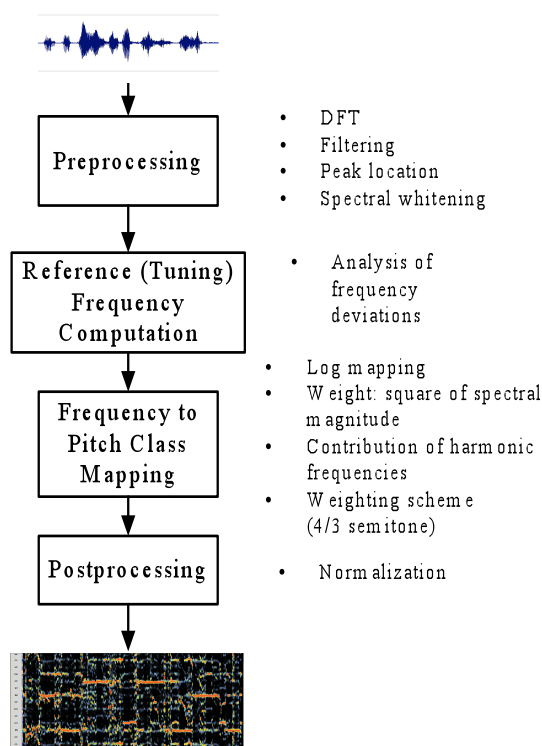


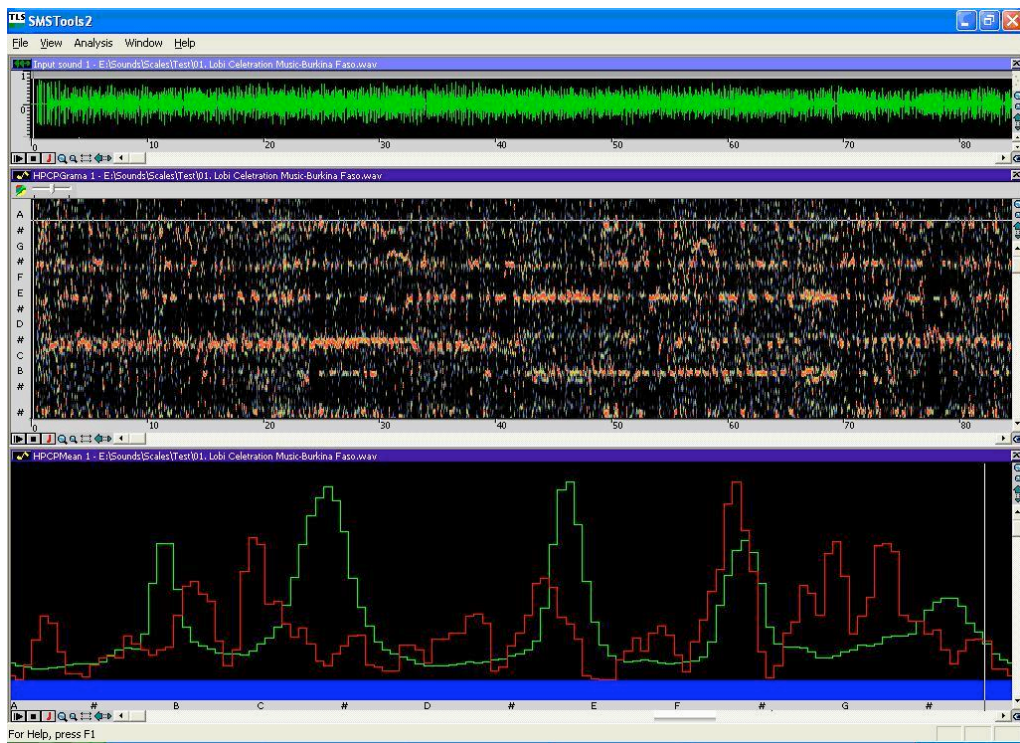
Fig. 3. Block diagram for HPCP computation.

We also compute a “transposed” version of the HPCP, called the THPCP, by ring-shifting the HPCP vector according to the position of the maximum value (*shift*):

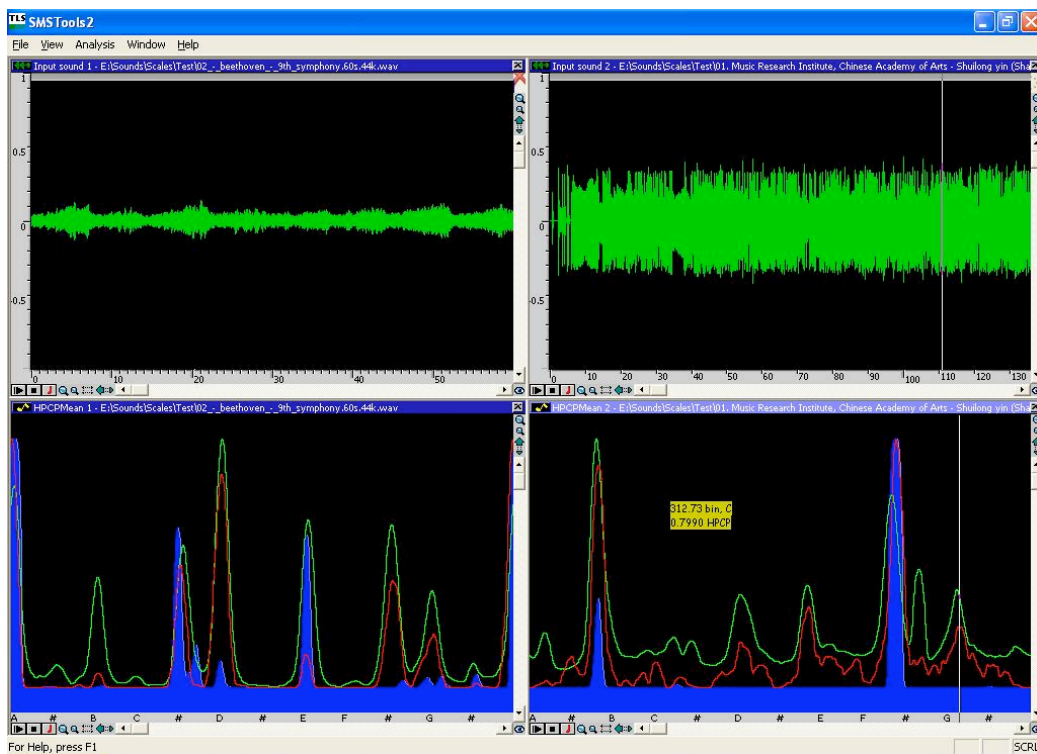
$$THCP[n] = HPCP[\text{mod}(n - \text{shift}, \text{size})] \quad n = 1, \dots, \text{size}.$$

This feature can be considered to be invariant to transposition, i.e. a piece which is transposed to a different key will have different HPCP but same THPCP values. In this sense, pieces are considered to be similar if they share the same tonal profile regardless of its absolute position.

Figure 4 shows an example of HPCP for an audio excerpt from Burkina Faso (labeled as *African*) including percussion and singing voice. We observed that the local maxima are not located on the exact positions of the tempered semitones. Figure 5 shows a comparison of high-resolution HPCP for a Western and non-Western musical excerpt. We also observe that the local maxima are not located on the exact positions of the tempered semitones for the Chinese piece, as it in fact happens for the Beethoven excerpt.



**Fig. 4.** Example of HPCP evolution for an African audio excerpt. Audio signal vs. time (top), high resolution HPCP vs. time (mid) and average HPCP (bottom, green line). The red line at the bottom shows a short-tem average of HPCP.



**Fig. 5.** HPCP for a Western and non-Western musical excerpt (Left: Beethoven 9<sup>th</sup> symphony. Right: Chinese folk instrumental piece). Audio signal vs. time (top), and global HPCP (bottom, green line). The red line at the bottom shows a short-tem average of HPCP, while the blue plot indicates an instantaneous value.



**Features derived from pitch class distributions**

We also compute two features from the ones described above. They are intended to distinguish between music composed using equal-tempered and non-equal tempered scales, as it is a relevant aspect to distinguish music from Western tradition.

**EQUAL-TEMPERED DEVIATION**

The equal-tempered deviation measures the deviation of the HPCP local maxima from equal-tempered bins. In order to compute this descriptor, we first extract a set of local maxima from the HPCP,  $\{k\}$ ,  $k=1..K$ , and we then compute their deviations from closest equal-tempered bins, weighted by their magnitude and normalized by the sum of peak magnitudes:

$$Etd = \frac{\sum_{k=1}^K HPCP[k] \cdot abs(k - et_k)}{\sum_{k=1}^K HPCP[k]}$$

where  $et_k$  represents the closest equal-tempered bin from HPCP bin  $k$ .

**NON-TEMPERED ENERGY RATIO**

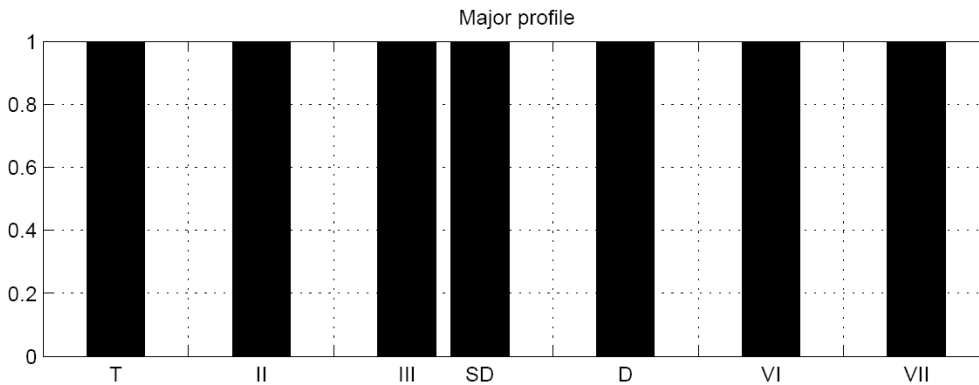
Non-tempered energy ratio represents the ratio between the amplitude of non-tempered HPCP bins and the total amplitude:

$$ER = 1 - \frac{\sum_{i=1}^{12} HPCP[pos_i]}{\sum_{i=1}^{size} HPCP[i]}$$

where  $size = 120$  (size of the HPCP vector) and  $pos_i$  are given by the HPCP positions related to the equal-tempered pitch classes.

**DIATONIC STRENGTH**

This descriptor represents the maximum correlation of the HPCP vector and a diatonic major profile ring-shifted in all possible positions. This diatonic major profile is represented in Figure 6.



**Fig. 6.** Diatonic major profile represented as the presence (1 or 0) of each of the 12 semitones of the chromatic scale.

We hypothesize that this correlation should be higher for a piece using a diatonic major scale, which is characteristic of Western music.

### Octave centroid

HPCP and related descriptors do not consider the absolute pitch height of the analyzed piece, as these descriptors map all the pitch class values to a single octave. This means that two pieces having the same pitch-class distribution but played in different octaves will have similar values for the extracted features. In order to take into account the octave location of the played piece, and to analyze its distribution over different musical pieces, we have defined a feature called *octave centroid*, which represents the geometry centre of the played pitches.

In order to estimate this descriptor, we first perform a spectral analysis and apply a multipitch estimation method based on Klapuri (2004). This method outputs an estimation of the predominant pitches found in the analyzed spectrum. A centroid feature is then computed from this representation on a frame basis. We finally consider different statistics of frame values (mean, median, standard deviation, inter-quartile range, kurtosis, skewness) as global descriptors for the analyzed piece.

### Roughness

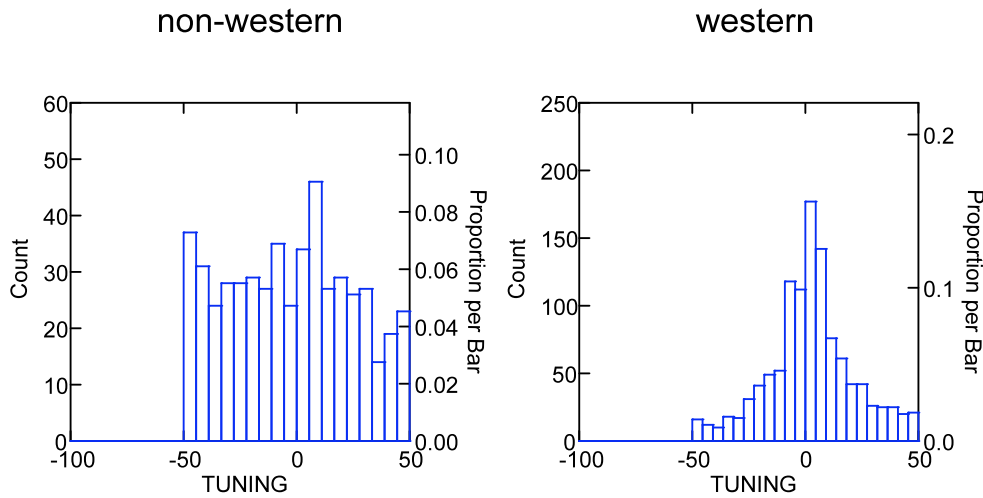
Complementary to pitch-class distribution and derived features, we also compute a roughness descriptor, as a measure of sensory dissonance. Roughness is a perceptual sensation arising from the presence of energy in very close frequencies, as it happens in the case that 2 instruments are less-than perfectly tuned. In order to experience roughness the frequency differences between partials has to be larger than 10 Hz (Zwicker and Fastl, 1999). Roughness can be typically experienced when listening to gamelan music, as the involved instruments are slightly mistuned on purpose (Tenzer, 1998).

We have used the roughness estimation model proposed in Vassilakis (2001, 2005). We compute a roughness value for each analysis frame, by summing the roughness of all pairs of components in the spectrum. We consider as the main frequency components of the spectrum those with spectral magnitudes higher than the 14% of the maximum spectral amplitude, as indicated in Vassilakis (2001, 2005). Then, a global roughness value is computed as the median of instantaneous values. We finally consider different statistics of frame values (mean, median, standard deviation, inter-quartile range, kurtosis, skewness) as global roughness descriptors for the analyzed piece.

## DISTRIBUTION OF FEATURES

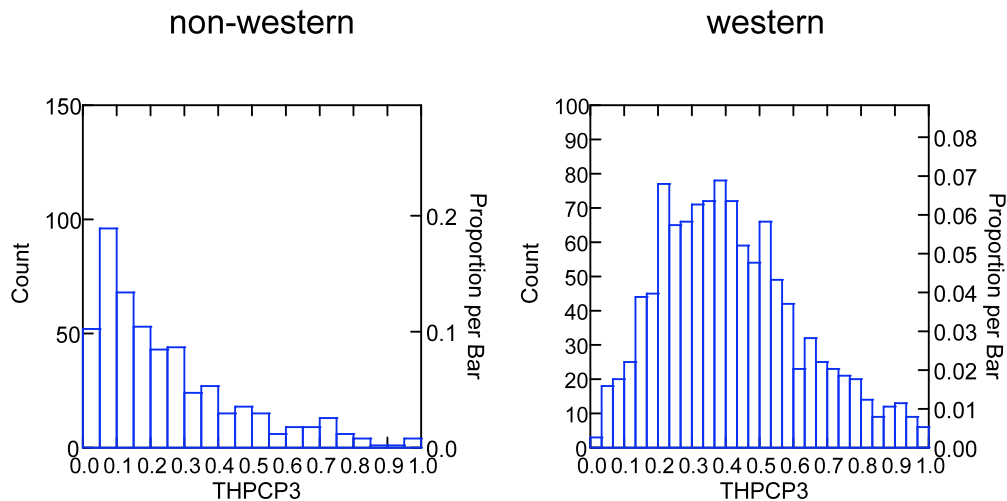
Once the set of features used to characterize the target collection has been defined and computed, we can study their distribution through the different types of music within the test collection, in order to have a preliminary idea of their usefulness for comparative analysis and classification. We present here some results on statistical analysis of the audio features for different group of pieces, in order to illustrate the differences between musical genres and origins.

Figure 7 shows the distribution of the tuning descriptor for Western and non-Western music. As expected, the distribution of tuning deviation with respect to 440 Hz is centered on 0 cents for Western music. On the other hand, it appears to be equal distributed between -50 and 50 cents for non-Western pieces. This is confirmed by a goodness-of-fit chi-square statistical test, where a p-value=0.061 indicates that the distribution of tuning frequencies for the non-Western pieces roughly follows a uniform distribution. On the other hand, the distribution for the Western pieces does not follow such a distribution.



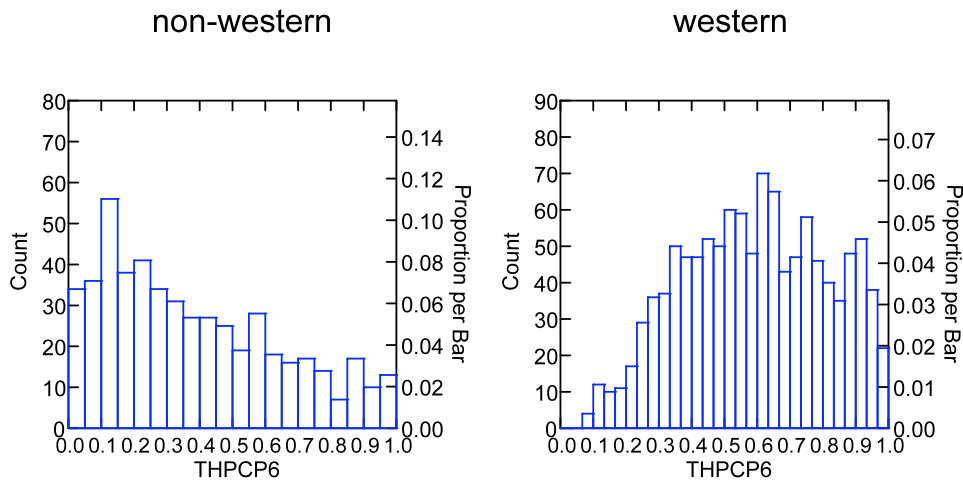
**Fig. 7.** Distribution of tuning deviation (in cents) for Western and non-Western music.

We also find some differences in the main transposed HPCP features, which represents the intensity of the different degrees of a diatonic major scale. For the main degrees, we find larger values for Western than for non-Western music. Figure 7 shows the distribution of THPCP3, which represents the intensity of the second degree of a diatonic scale. According to the distribution, it appears to be lower for non-Western music than for Western music, and differently distributed. In fact, a goodness-of-fit chi-square statistical test yields a p-value=0.132, indicating the correspondence between THPCP3 distribution of non-Western pieces and a Gompertz distribution (with parameters  $b = 3.2$  and  $c = 2.62$ ). The same statistical test yields a p-value=0 for the same distribution considering Western pieces.



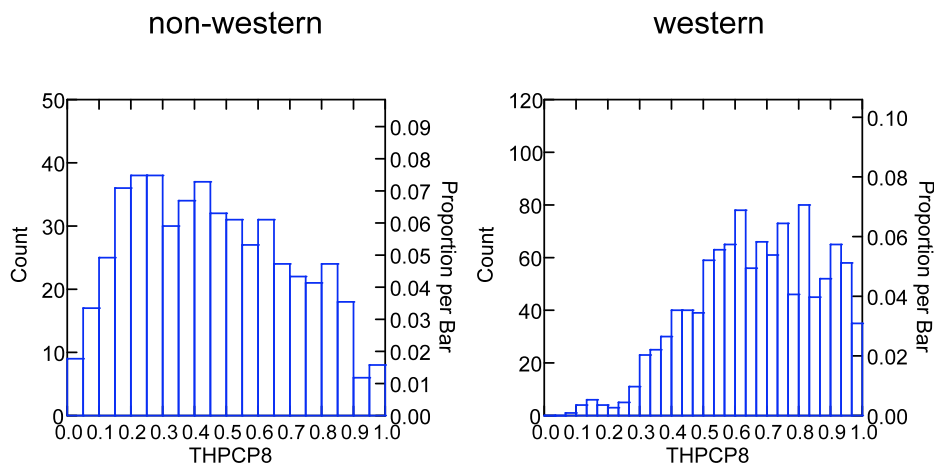
**Fig. 8.** Distribution of THPCP3 for Western and non-Western music.

The THPCP6 feature represents the intensity of the fourth (sub-dominant) degree of a diatonic scale. This feature also shows lower values for non-Western music than for Western music and they are differently distributed, as shown in figure 9.



**Fig. 9.** Distribution of THPCP6 for Western and non-Western music.

THPCP8 represents the intensity of the fifth (dominant) degree of a diatonic scale. As for the previous descriptors, it seems to be lower for non-Western music than for Western music and differently distributed, as shown figure 10. In fact, a goodness-of-fit chi-square statistical test yields a p-value=0.652, indicating the correspondence between THPCP8 distribution from non-Western pieces and a Logit Normal distribution (with parameters  $\mu=-0.206171$   $\sigma=1.293628$ ). The same statistical test yields a p-value=0 for the same distribution considering Western pieces.



**Fig. 10.** Distribution of THPCP8 for Western and non-Western music.

Regarding descriptors derived from HPCP, the equal-tempered deviation, representing the deviation from an equal-tempered scale, also appears to be lower for Western than for non-Western music, as shown in figure 11. In fact, a goodness-of-fit chi-square statistical test yields a p-value=0.652, indicating the correspondence between the equal-tempered deviation distribution from non-Western pieces and a Weibull distribution (scale=0.185 and shape=2.156). The same statistical test yields a p-value=0 for the same distribution considering Western pieces.

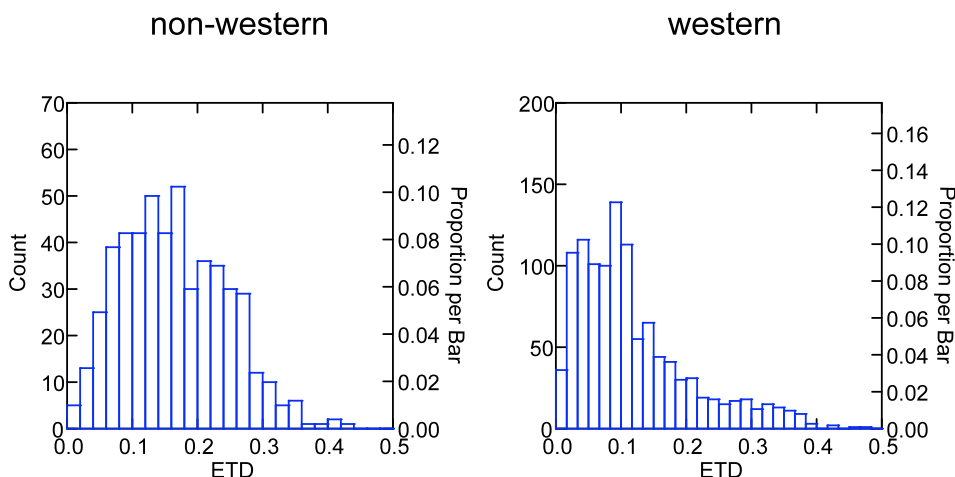


Fig. 11. Distribution of equal-tempered deviation for Western and non-Western music.

### AUTOMATIC CLASSIFICATION

After this preliminary statistical analysis of the computed features, we present here some results of automatic classification based on the extracted descriptors. Our goal here is to have a classifier that automatically assigns the label “Western” or “non-Western” to any audio file that is input and analyzed according to the procedure explained in the previous sections.

We have used different machine learning techniques implemented in the WEKA machine learning software, a collection of classification algorithms for machine learning tasks (Witten & Frank, 2005b). Weka contains tools for data pre-processing, classification, regression, clustering, association rules and visualization. The algorithms are applied directly to our dataset.

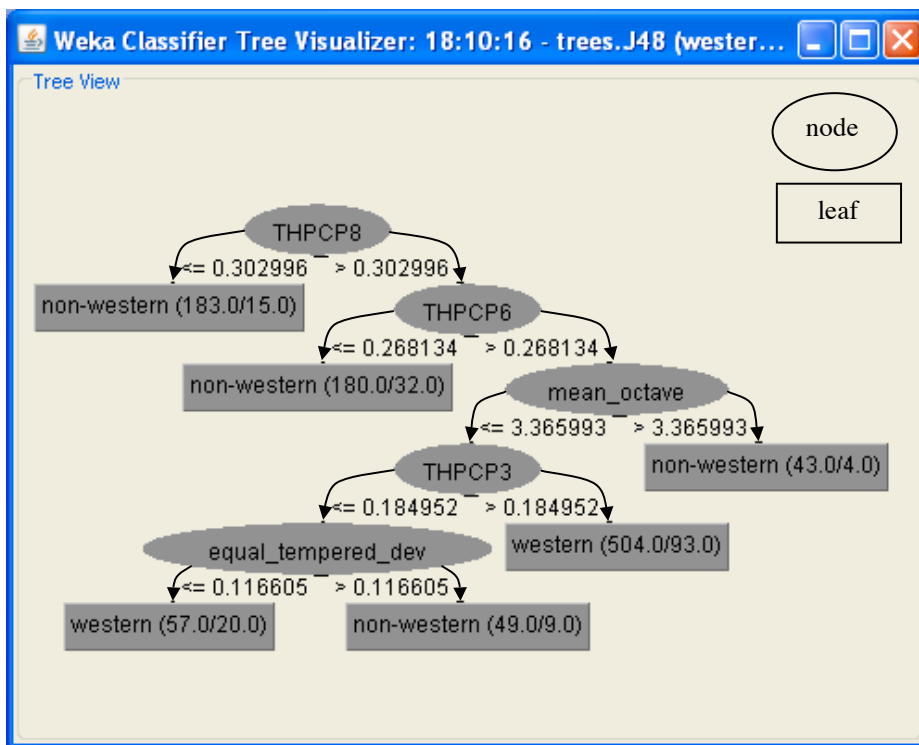
#### Evaluation results using different machine learning techniques

We adopt an evaluation procedure based on 10-fold cross-validation over equally-distributed classes. In 10-fold cross-validation, the data is divided into 10 subsets of approximately equal number of examples, and the different machine learning algorithms are trained 10 times, each time leaving out one of the subsets from training, but using only the omitted subset for accuracy estimation, this way we test the learnt model using previously “unseen” examples. We do this train-test cycle ten times in order to provide a better estimate of the generalization error. In other words, we try to get a good estimation of the errors the system will yield when classifying examples that do not belong to our current collection.

We have approached several classification methods but, for the sake of clarity and conciseness we only present two of them, which are explained in detail in (Witten & Frank, 2005a). One of these approaches (decision trees) provides a clearly understandable output that has allowed identifying the most relevant features for classification. In addition, decision trees are easy to implement as a collection of “if-then” rules in any programming environment. The other approach we have explored (support vector machine) is considered as one of the best-performing learning algorithms currently available.

- **Decision trees** are massively used for different machine learning and classification tasks. One of the main reasons for their acceptance lies in the fact that their output is a model that can be interpreted as a series of {if a descriptor has a value bigger or smaller than x then classify the observation as C} clauses. Decision trees are constructed top-down, beginning with the feature that seems to be the most informative one, that is, the one that maximally reduces entropy. Branches are then created from each one of the different values of this feature. The training examples are sorted to the appropriate descendant node, and the entire process is then repeated recursively using the examples of one of the descendant nodes, then those of the other. An in-depth treatment of decision trees can be found in (Mitchell, 1997). As shown in Figure 12, which depicts the decision tree computed to model our data, the test at a node compares the descriptor with a constant value. Leaf nodes give a classification that

applies to all instances that reach the leaf. There are different algorithms to build decision trees. The one we have used, called *J4.8* in Weka is an implementation of the version 8 of the so-called C4.5 (Quinlan, 1993; Witten & Frank, 2005a, p. 189-200), which is probably the most often decision tree used in the scientific community. We have also tested different values for the parameter *minObj*, which specifies the minimum number of objects (or instances) allowed at a leaf. This allows getting very compact trees without sacrificing precision as figure 12 illustrates.



**Fig. 12.** Example of a Decision tree (J48 with a minimum of 40 examples per leaf) with 6 leaves. Parentheses indicate the number of correctly/incorrectly classified files for each branch. The overall classification accuracy for this tree is provided in Table 1.

- Support Vector Machines (SVM)** are classifiers based on statistical learning theory (Vapnik, 1998). The basic training principle underlying SVMs is finding the optimal linear hyperplane that separates two classes, such that the expected classification error for unseen test samples is minimized (i.e., they look for good generalization performance). This hyperplane can be delimited by a subset of the available instances, which define the “support vectors” for it. Based on this principle, a SVM uses a systematic approach to find a linear function with the lowest complexity. For linearly non-separable data, SVMs can (non-linearly) map the input to a high dimensional feature space where a linear hyperplane can be found. This mapping is done by means of a so-called *kernel* function. Although there is no guarantee that a linear solution will always exist in the high dimensional space, in practice it is quite feasible to construct a working solution. In other words, it can be said that training a SVM is equivalent to solving a quadratic programming with linear constraints and as many variables as data points. Weka implements support vector machines using the SMO (Sequential minimal optimization) algorithm (Witten & Frank, 2005a, p. 214-235; Platt, 1998). It is advisable to tune certain parameters of this algorithm, as its complexity parameter and the exponent for the used polynomial kernel, in order to decrease its classification error.

For the Western versus non-Western categories, and using these different techniques, we achieve the classification results summarized in Table 1. F-measure is a common measure to evaluate the performance of information retrieval systems, and it is defined as the weighted harmonic mean of precision and recall:

$$F\text{-measure} = 2 \cdot (\text{precision} \cdot \text{recall}) / (\text{precision} + \text{recall})$$

where precision is the fraction of the retrieved instances that belong to the correct category and recall is the fraction of the documents that belong to the correct category which are successfully retrieved.

Method		Global accuracy	Accuracy per class	
Classification algorithm	WEKA Parameters	Correctly Classified Instances	F-Measure Western	F-Measure non-Western
Decision Trees (J48)	30 minObj	80.41 %	0.808	0.8
	40 minObj	<b>82.38 %</b>	0.827	0.82
	50 minObj	79.63 %	0.795	0.797
Support Vector Machine (Binary SMO Classifier)	Complexity = 1, Exponent = 1.5	86.12 %	0.865	0.857
	Complexity = 1.5, Exponent = 1.5	<b>86.51 %</b>	0.869	0.861

**Table 1.** Classification results for different machine learning techniques provided by WEKA.

As a general conclusion, we observed that the classification accuracy is higher than 80% for all the machine learning techniques used. It could be argued that the Western-non Western distinction that we are aimed at is an artificial and biased one. If that was the case, an analysis that would not superimpose those two categories as an “a priori” could yield different groupings of the data. Cluster analysis allows for that kind of unsupervised, emergent type of data arrangement. K-means clustering is one of the most usual clustering techniques (Witten & Frank, 2005a), and it tries to group data into homogeneous clusters, and, at the same time, to separate heterogeneous data into different clusters. The homogeneity criterion is defined by means of a Euclidean distance, which the algorithm tries to minimize for the examples that are clustered together and maximized among different clusters.

In our experiment, we have clustered the data asking for the algorithm to find 2 clusters. We performed 10 runs of the algorithm and cross-tabulated the cluster assignment against the Western/non-Western distinction, observing that the error varied from 29.677% to 29.9817%, which indicates an extremely robust solution: the two clusters found in a non-supervised way coincide, to a large extent, with the two categories used in the supervised experiments. In addition, clustering solutions using more than 2 groups yielded larger errors than the 2-clusters solution. It could, then, be the case that the distinction is not so artificial, and that it emerges when we consider music according to the tonality-related descriptors that we have computed.

## Feature selection

The classifiers used in the experiments presented above are designed to detect the most appropriate descriptors and even the most appropriate instances for optimizing their decisions. However, there are some automatic methods that specifically give some hints on the usefulness of the available features.

We have tested an attribute evaluation method for attribute selection, correlation-based feature selection (CFS) (Hall, 2000). This algorithm selects a near-optimal subset of features that have minimal correlation between them, and maximal correlation with the to-be-predicted classes. In the set of descriptors selected by this algorithm we observed that the most relevant descriptors are THPCP3, THPCP8, THPCP10, tuning, equal tempered deviation, roughness (mainly its median and standard deviation along the considered excerpt) and octave centroid (average along the audio excerpt).

The relevance of these features has already been noticed when performing an analysis of their value distributions and they also coincide with the descriptors found on the generated decision trees, as shown in the previous section.

### Classification accuracy for different musical genres and traditions

It is very informative to study the distribution of classification errors, in order to have some insights about the limitations of the method. For instance, we observed in Table 1 that the accuracy (F-measure) for Western and non-Western categories has no significant differences.

In order to study the accuracy distribution over musical genres, we have built the decision tree shown in Figure 12, which provides a classification accuracy of 82.38% over the music collection under study. Figure 13 shows the distribution of correct classification for different Western musical genres and non-Western traditions (grouped by region).

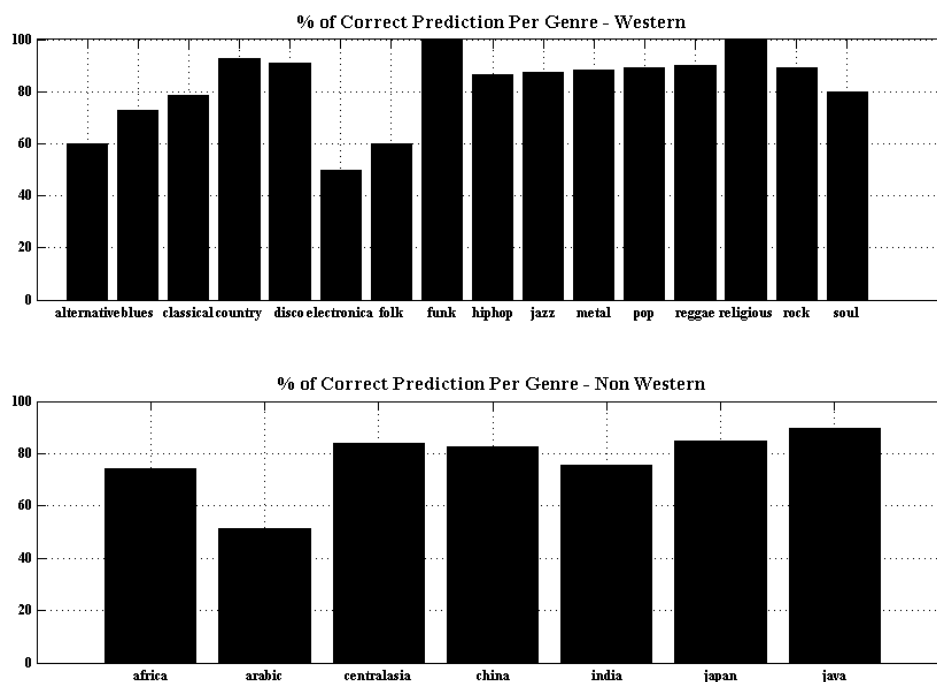


Fig. 13. Percentage of correctly classified instances distributed among genres using the Decision Tree from Figure 12. The classification accuracy for this tree is provided in Table 1.

We observed that *electronica* is the most misclassified Western style, and it is usually labeled as non-Western. We can justify this by the fact that the excerpts under electronic music usually include non-quantized pitched sounds, as well as non-pitched sounds that can increase the values of features such as equal tempered deviation or tuning, which have been found to be high for non-Western music. This is also the case for the few items under the *alternative* category. In the case of the 10 *folk* examples, 40% of them are misclassified due to a low value of THPCP6, related to the relative intensity of the fourth degree of the scale.

Regarding non-Western material, *Arabic* music is sometimes labeled as Western. This can be due to the tonal similarity between some of the audio excerpts under this category and the Western musical tradition, including some tempered instruments and scale degrees. An analysis of the misclassified instances revealed high values for the descriptors THPCP8 (related to the fifth), THPCP6 (related to the fourth) and THPCP3 (related to the second) in 81.82% of the misclassified instances, and high values for THPCP8 and THPCP6 together with small equal tempered deviation for the rest.

### Classification accuracy for an independent test set

In order to test the generalization power of the classification algorithm, we have performed an evaluation using an independent test set from the collection used for training the models. The chosen independent set is the NASA Voyager Golden Record [1]. NASA placed aboard the Voyager spacecraft a time capsule intended to represent the history of our world to extraterrestrials. This capsule contains a record with sound and images of Earth, selected by a committee chaired by Carl Sagan of Cornell University and including 27



musical pieces from Eastern and Western classics and a variety of ethnic music (Sagan, 1984). These 27 pieces were manually classified by us into 12 Western and 17 non-Western pieces, and then analyzed in order to obtain the set of descriptors from the 30 first seconds of each piece. These data were then used as an independent test set for the selected classifiers obtaining the classification results shown in Table 2.

Method		Global accuracy	Accuracy per class	
Classification algorithm	Parameters	Correctly Classified Instances	F-Measure Western	F-Measure non-Western
Decision Trees (J48)	30 minObj	70.34 %	0.667	0.733
	40 minObj	<b>85.18 %</b>	0.818	0.875
	50 minObj	<b>85.18 %</b>	0.818	0.875
Binary SMO Classifier	Complexity = 1, Exponent = 1	81.48 %	0.737	0.857
	Complexity = 1, Exponent = 1.5	74.07 %	0.632	0.8
	Complexity = 1.5, Exponent = 1.5	74.07 %	0.632	0.8

**Table 2.** Classification results for different machine learning techniques provided by WEKA, using an independent test set from the Voyager Golden Record.

We observed that 85.18% of the pieces were correctly classified by the system, even if the obtained performance slightly varied for the two classification algorithms. The best results were obtained by using decision trees with different configuration parameters. The classifier reaches nearly the same performance with these “unseen” files than in the train-test cycle, which demonstrates the generalization ability of the proposed models.

## CONCLUSIONS AND FUTURE WORK

In this study we provided an empirical approach to the comparative analysis of music audio recordings, focusing on tonal features and a music collection from different traditions and musical styles. We presented some encouraging results obtained when trying to automatically distinguish or classify music from Western and non-Western traditions by means of automatically audio feature extraction and data mining techniques. We obtained a high rate of classification accuracy of 80% for a music collection of 1500 pieces from different musical traditions using a restricted set of tonal features. From this, we can argue that it can be possible to automatically classify music into western and non-western by just analyzing audio data.

We are aware that there are larger issues involved in the determination of musical genres. We are also aware of the limitations of the concept of Western as opposed to non-Western music. Ideally we should be able to define and formalize stylistic features proper to different kinds of music or “stylistic areas” and approach genres not just geographically but as a set of traits and then refine our descriptors accordingly.

Future work will then be devoted to different issues. One important aspect that we would like to achieve is to contrast and complement this group of descriptions from an ethnomusicology perspective, analyzing in detail some of the used excerpts. We will also investigate the main variations inside Western and non-Western styles by comparing in details the different musical genres and musical traditions. We also plan to analyze how automatically-extracted audio features related to timbre and rhythmic aspects (which were out of our scope in this paper) can improve the classification and complement the current feature set.

The present work shows that automatic audio description tools, together with data mining techniques can help to characterize huge music collections and complement musicological manual analyses. It also confirms that tonal features extracted from audio data are representative of the pitch class

distribution, scale, gamut and tuning system of the analyzed piece, and that they provide means of characterizing different traditions and styles. We believe that audio description tools have a great potential to assist in ethnomusicological research and we hope that our work will contribute to the understanding of the world's musical heritage by means of computational modeling.

## ACKNOWLEDGMENTS

The authors would like to thank Dr. Ricardo Canzio, from the Graduate Institute of Musicology, National Taiwan University, for his expertise about ethnomusicology and his assistance and review of this project. Special thank also go to Professor Paul Poletti, from the *Escola Superior de Música de Catalunya* for his knowledge and recommendations about scales and tuning systems in different musical traditions, and to Dr. Ramon Pelinski for his insightful comments related to the ethnomusicological issues involved in our research.

## NOTES

[1] <http://voyager.jpl.nasa.gov/spacecraft/goldenrec.html>

## REFERENCES

- Burns, E. M. (1998). *Intervals, scales and tuning*. In *The Psychology of Music*, 2<sup>nd</sup> Edition, edited by Diana Deutsch, pp. 215-264. New York: Academic Press. ISBN 0-12-213564-4.
- Carterette, E. C., & Kendall, R. A. (1994). *On the tuning and stretched octave of Javanese gamelan*. *Leonardo Music Journal*, Vol 4. pp. 59-68.
- Drabkin, W. (2008) *Scale*. Grove Music Online ed. L. Macy (Accessed [31 January 2008]), <http://www.grovemusic.com>
- Gómez, E. (2006). *Tonal description of music audio signals*. PhD dissertation, Universitat Pompeu Fabra. <http://mtg.upf.edu/~egomez/thesis>
- Gouyon, F., & Dixon, S. (2005). *A review of automatic rhythm description systems*. *Computer Music Journal* Vol. 29, No. 1, pp. 34-54.
- Hall, M. A. (2000). *Correlation-based feature selection for discrete and numeric class machine learning*. In proceedings of the Seventeenth International Conference on Machine Learning,
- Holzapfel, A., & Stylianou, Y. (2007). *A Statistical Approach To Musical Genre Classification Using Non-Negative Matrix Factorization*, Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vol. 2(15-20), April 2007, pp. II-693 - II-696.
- Klapuri, A. (2004). *Signal processing methods for the automatic transcription of music*, Ph.D. thesis, Tampere University of Technology, Finland.
- Lesaffre, M., Leman, M., De Baets, B., & Martens, J.-P. (2004). *Methodological considerations concerning manual annotation of musical audio in function of algorithm development*, Proceedings of the International Conference on Music Information Retrieval (ISMIR), Barcelona, Spain, October 9-14.
- Merriam, A. P. (1959). *Characteristics of African Music*. *Journal of the International Folk Music Council*, Vol. 11, pp. 13-19.
- Mitchell, T. M. (1997). *Machine learning*. Boston, MA: McGraw-Hill.

- Piggott, F. T. (1891-1892). *The Music of Japan*. Proceedings of the Musical Association, 18<sup>th</sup> Sess., pp. 103-120.
- Platt, J. (1998). *Fast Training of Support Vector Machines using Sequential Minimal Optimization*, Advances in Kernel Methods - Support Vector Learning, B. Schoelkopf, C. Burges, and A. Smola, eds., MIT Press.
- Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers.
- Rentfrow, P. J., & Godsling, S. D. (2003). *The Do Re Mi's of Everyday Life: The Structure and Personality Correlates of Music Preferences*. Journal of Pers. Soc. Psychology, Vol. 84, No. 6, pp. 1236-1254.
- Sagan, C (1984). *Murmurs of Earth: The Voyager Interstellar Record*. Ballantine Books.
- Tenzer, M. (1998). *Balinese Music*. Periplus Editions: Singapore.
- Toivianen, P., & Eerola, T. (2006). *Visualization in comparative music research*. In A. Rizzi & M Vichi (Eds.), COMPSTAT 2006 - Proceedings in Computational Statistics. Heidelberg: Physica-Verlag, pp. 209-221.
- Tzanetakis G., Kapur, A., Andrew Schloss, W., & Wright, M. (2007). *Computational Ethnomusicology*, Journal of Interdisciplinary Music Studies, Vol. 1, No. 2, pp. 1-24.
- Tzanetakis, G., Essl, G., & Cook, P.R. (2001). *Automatic Musical Genre Classification of Audio Signals*, Proceedings of International Symposium on Music Information Retrieval (ISMIR), Bloomington, Indiana.
- Tzanetakis, G., & Cook, P. (2002). *Musical Genre Classification of Audio Signals*, IEEE Transactions on Speech and Audio Processing , Vol. 10, No. 5, July 2002.
- Vapnik, V.N. (1998). *Statistical learning theory*. New York: Wiley
- Vassilakis, P. N. (2001). *Perceptual and Physical Properties of Amplitude Fluctuation and their Musical Significance*. Doctoral Dissertation. Los Angeles: University of California, Los Angeles; Systematic Musicology.
- Vassilakis, P. N. (2005). *Auditory roughness as a means of musical expression*, Selected Reports in Ethnomusicology 12 (Perspectives in Systematic Musicology), pp. 119-144.
- VV, A A. (1973). *The scales of some African Instruments*. International Library of African Music, Sound of Africa Series – LP Records, pp. 91-107 <http://anaphoria.com/depos.html>
- Witten, I. H., & Frank, E. (2005a). *Data Mining. Practical Machine Learning Tools and Techniques*, Second Edition, Elsevier, San Francisco, CA, USA.
- Witten, I. H., & Frank, E. (2005b). *Weka 3: Data Mining Software in Java (Version 3.4)* [Computer Software]. Available from <http://www.cs.waikato.ac.nz/ml/weka/>
- Zwicker, E., & Fastl, H. (1999). *Psychoacoustics: Facts and models*. Berlin: Springer.

Bogdanov, D., Wack, N., Gómez, E., Gulati S., **Herrera, P.**, Mayor, O., Roma, G., Salamon, J., Zapata, J. & Serra, X. (2014). “ESSENTIA: an open source library for audio analysis”. ACM SIGMM Records. 6(1).

<http://records.mlab.no/2014/03/20/essentia-an-open-source-library-for-audio-analysis/>

ISSN 1947-4598

3. **81** - Real-time bag of words, approximately  
J. R. R. Uijlings, A. W. M. Smeulders, R. J. H. Scha  
<http://dl.acm.org/citation.cfm?id=1646405>
4. **57** - Dense sampling and fast encoding for 3D model retrieval using bag-of-visual features  
Hideki Nakayama, Tatsuya Harada, Yasuo Kuniyoshi  
<http://dl.acm.org/citation.cfm?id=1646419>
5. **46** - Multilayer pLSA for multimodal image retrieval  
Rainer Lienhart, Stefan Romberg, Eva Hörster  
<http://dl.acm.org/citation.cfm?id=1646408>
3. **119** - Learning tag relevance by neighbor voting for social image retrieval  
Xirong Li, Cees G.M. Snoek, Marcel Worring  
<http://dl.acm.org/citation.cfm?id=1460126>
4. **58** - Spirittagger: a geo-aware tag suggestion tool mined from flickr  
Emily Moxley, Jim Kleban, B. S. Manjunath  
<http://dl.acm.org/citation.cfm?id=1460102>
5. **42** - Content-based mood classification for photos and music: a generic multi-modal classification framework and evaluation approach  
Peter Dunker, Stefanie Nowak, André Begau, Cornelia Lanz  
<http://dl.acm.org/citation.cfm?id=1460114>

#### CIVR 2010

1. **43** - Signature Quadratic Form Distance  
Christian Beecks, Merih Seran Uysal, Thomas Seidl  
<http://dl.acm.org/citation.cfm?id=1816105>
2. **41** - Feature detector and descriptor evaluation in human action recognition  
Ling Shao, Riccardo Mattivi  
<http://dl.acm.org/citation.cfm?id=1816111>
3. **38** - Unsupervised multi-feature tag relevance learning for social image retrieval  
Xirong Li, Cees G. M. Snoek, Marcel Worring  
<http://dl.acm.org/citation.cfm?id=1816044>
4. **29** - Co-reranking by mutual reinforcement for image search  
Ting Yao, Tao Mei, Chong-Wah Ngo  
<http://dl.acm.org/citation.cfm?id=1816048>
5. Two papers were tied for 5th place in citations:
  - **20** - On the sampling of web images for learning visual concept classifiers  
Shiai Zhu, Gang Wang, Chong-Wah Ngo, Yu-Gang Jiang  
<http://dl.acm.org/citation.cfm?id=1816051>
  - **20** - Plant species identification using leaf image retrieval  
Carlos Caballero, M. Carmen Aranda  
<http://dl.acm.org/citation.cfm?id=1816089>

#### MIR 2010

1. **285** - The MIR flickr retrieval evaluation  
Mark J. Huiskes, Michael S. Lew  
<http://dl.acm.org/citation.cfm?id=1460104>
2. **203** - Outdoors augmented reality on mobile phone using loxel-based visual feature organization  
Gabriel Takacs, Vijay Chandrasekhar, Natasha Gelfand, Yingen Xiong, Wei-Chao Chen, Thanos Bismpiagiannis, Radek Grzeszczuk, Kari Pulli, Bernd Girod  
<http://dl.acm.org/citation.cfm?id=1460165>

#### MIR 2010

1. **82** - New trends and ideas in visual concept detection: the MIR flickr retrieval evaluation initiative  
Mark J. Huiskes, Bart Thomee, Michael S. Lew  
<http://dl.acm.org/citation.cfm?id=1743475>
2. **78** - How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation  
Stefanie Nowak, Stefan Rürger  
<http://dl.acm.org/citation.cfm?id=1743478>
3. **45** - Exploring automatic music annotation with "acoustically-objective" tags  
Derek Tingle, Youngmoo E. Kim, Douglas Turnbull  
<http://dl.acm.org/citation.cfm?id=1743400>
4. **39** - Feature selection for content-based, time-varying musical emotion regression  
Erik M. Schmidt, Douglas Turnbull, Youngmoo E. Kim  
<http://dl.acm.org/citation.cfm?id=1743431>
5. **34** - ACQUINE: aesthetic quality inference engine – real-time automatic rating of photo aesthetics  
Ritendra Datta, James Z. Wang  
<http://dl.acm.org/citation.cfm?id=1743457>

## ESSENTIA: an open source library for audio analysis

Over the last decade, audio analysis has become a field of active research in academic and engineering worlds. It refers to the extraction of information and meaning from audio signals for analysis, classification, storage, retrieval, and synthesis, among other tasks. Related research topics challenge understanding and

modeling of sound and music, and develop methods and technologies that can be used to process audio in order to extract acoustically and musically relevant data and make use of this information. Audio analysis techniques are instrumental in the development of new audio-related products and services, because these techniques allow novel ways of interaction with sound and music.



**Essentia** is an open-source C++ library for audio analysis and audio-based music information retrieval released under the **Affero GPLv3 license** (also available under proprietary license upon request). It contains an extensive collection of reusable algorithms which implement audio input/output functionality, standard digital signal processing blocks, statistical characterization of data, and a large set of spectral, temporal, tonal and high-level music descriptors that can be computed from audio. In addition, Essentia can be complemented with **Gaia**, a C++ library with python bindings which allows searching in a descriptor space using different similarity measures and classifying the results of audio analysis (same license terms apply). Gaia can be used to generate classification models that Essentia can use to compute high-level description of music.

Essentia is not a framework, but rather a collection of algorithms wrapped in a library. It doesn't enforce common high-level logic for descriptor computation (so you aren't locked into a certain way of doing things). It rather focuses on the robustness, performance and optimality of the provided algorithms, as well as ease of use. The flow of the analysis is decided and implemented by the user, while Essentia is taking care of the implementation details of the algorithms being used. A number of examples are provided with the library, however they should not be considered as the only correct way of doing things.

The library includes **Python bindings** as well as a number of predefined executable extractors for the available music descriptors, which facilitates its use for fast prototyping and allows setting up research experiments very rapidly. The extractors cover a number of common use-cases for researchers, for example, computing all available music descriptors for an audio track, extracting only spectral, rhythmic, or tonal descriptors, computing predominant melody and beat positions, and returning the results in yaml/json data formats. Furthermore, it includes a **Vamp plugin** to be used for visualization of music descriptors using hosts such as Sonic Visualiser.

The library is **cross-platform** and supports Linux, Mac OS X and Windows systems. Essentia is designed with a focus on the robustness of the provided music descriptors and is optimized in terms of the computational cost of the algorithms. The provided functionality, specifically the music descriptors included out-of-the-box and signal processing algorithms, is **easily expandable** and allows for both research experiments and development of large-scale industrial applications.

Essentia has been in development for more than 7 years incorporating the work of more than 20 researchers and developers through its history. The 2.0 version marked the first release to be publicly available as free software released under AGPLv3.

## Algorithms

Essentia currently features the following algorithms (among others):

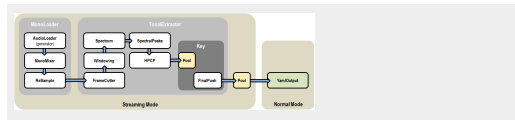
- **Audio file input/output:** ability to read and write nearly all audio file formats (wav, mp3, ogg, flac, etc.)
- **Standard signal processing blocks:** FFT, DCT, frame cutter, windowing, envelope, smoothing
- **Filters (FIR & IIR):** low/high/band pass, band reject, DC removal, equal loudness
- **Statistical descriptors:** median, mean, variance, power means, raw and central moments, spread, kurtosis, skewness, flatness
- **Time-domain descriptors:** duration, loudness, LARM, Leq, Vickers' loudness, zero-crossing-rate, log attack time and other signal envelope descriptors
- **Spectral descriptors:** Bark/Mel/ERB bands, MFCC, GFCC, LPC, spectral peaks, complexity, rolloff, contrast, HFC, inharmonicity and dissonance
- **Tonal descriptors:** Pitch salience function, predominant melody and pitch, HPCP (chroma) related features, chords, key and scale, tuning frequency
- **Rhythm descriptors:** beat detection, BPM, onset detection, rhythm transform, beat loudness
- **Other high-level descriptors:** danceability, dynamic complexity, audio segmentation, semantic annotations based on SVM classifiers

The complete list of algorithms is available online in the official documentation.

## Architecture

The main purpose of Essentia is to serve as a library of signal-processing blocks. As such, it is intended to provide as many algorithms as possible, while trying to be as little intrusive as possible. Each processing block is called an Algorithm, and it has three different types

of attributes: inputs, outputs and parameters. Algorithms can be combined into more complex ones, which are also instances of the base Algorithm class and behave in the same way. An example of such a composite algorithm is presented in the figure below. It shows a composite tonal key/scale extractor, which combines the algorithms for frame cutting, windowing, spectrum computation, spectral peaks detection, chroma features (HPCP) computation and finally the algorithm for key/scale estimation from the HPCP (itself a composite algorithm).



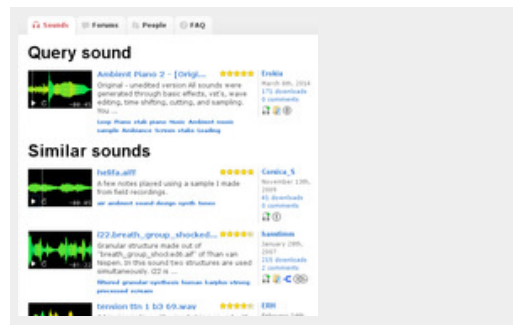
The algorithms can be used in two different modes: standard and streaming. The **standard mode** is imperative while the streaming mode is declarative. The standard mode requires to specifying the inputs and outputs for each algorithm and calling their processing function explicitly. If the user wants to run a network of connected algorithms, he/she will need to manually run each algorithm. The advantage of this mode is that it allows very rapid prototyping (especially when the python bindings are coupled with a scientific environment in python, such as ipython, numpy, and matplotlib).

The **streaming mode**, on the other hand, allows to define a network of connected algorithms, and then an internal scheduler takes care of passing data between the algorithms inputs and outputs and calling the algorithms in the appropriate order. The scheduler available in Essentia is optimized for analysis tasks, and does not take into account the latency of the network. For real-time applications, one could easily replace this scheduler with another one that favors latency over throughput. The advantage of this mode is that it results in simpler and safer code (as the user only needs to create algorithms and connect them, there is no room for him to make mistakes in the execution order of the algorithms), and in lower memory consumption in general, as the data is streamed through the network instead of being loaded entirely in memory (which is the usual case when working with the standard mode). Even though most of the algorithms are available for both the standard and streaming mode, the code that implements them is not duplicated as either the streaming version of an algorithm is deduced/wrapped from its standard implementation, or vice versa.

## Applications

Essentia has served in a large number of research activities conducted at Music Technology Group since

2006. It has been used for music classification, semantic autotagging, music similarity and recommendation, visualization and interaction with music, sound indexing, musical instruments detection, cover detection, beat detection, and acoustic analysis of stimuli for neuroimaging studies. Essentia and Gaia have been used extensively in a number of research projects and industrial applications. As an example, both libraries are employed for large-scale indexing and content-based search of sound recordings within **Freesound**, a popular repository of Creative Commons licensed audio samples. In particular, Freesound uses audio based similarity to recommend sounds similar to user queries. Dunya is a web-based software application using Essentia that lets users interact with an audio music collection through the use of musical concepts that are derived from a specific musical culture, in this case Carnatic music.



## Examples

Essentia can be easily used via its python bindings. Below is a quick illustration of Essentia's possibilities for example on detecting beat positions of music track and its predominant melody in a few lines of python code using the standard mode:

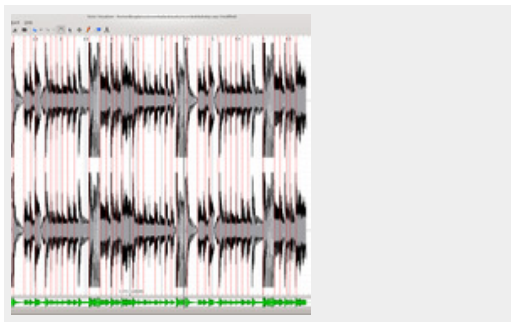
```
from essentia.standard import *;
audio = MonoLoader(filename =
'audio.mp3')(); beats, bconfidence =
```

```
BeatTrackerMultiFeature()(audio); print
beats; audio = EqualLoudness()
(audio); melody, mconfidence
= PredominantMelody(guessUnvoiced=True,
frameSize=2048, hopSize=128)(audio);
print melody
```

Another python example for computation of MFCC features using the streaming mode:

```
from essentia.streaming import * loader
= MonoLoader(filename = 'audio.mp3')
frameCutter = FrameCutter(frameSize =
1024, hopSize = 512) w = Windowing(type
= 'hann') spectrum = Spectrum() mfcc =
MFCC() pool = essentia.Pool() # connect
all algorithms into a network loader.audio
>> frameCutter.signal frameCutter.frame
>> w.frame >> spectrum.frame
spectrum.spectrum >> mfcc.spectrum
mfcc.mfcc >> (pool, 'mfcc') mfcc.bands
>> (pool, 'mfcc_bands') # compute network
essentia.run(loader) print pool['mfcc']
print pool['mfcc_bands']
```

Vamp plugin provided with Essentia allows to use many of its algorithms via the graphical interface of Sonic Visualiser. In this example, positions of onsets are computed for a music piece (marked in red):



An interested reader is referred to the documentation online for more example applications built on top of Essentia.

## Getting Essentia

The detailed information about Essentia is available online on the official web page: <http://essentia.upf.edu>. It contains the complete documentation for the project, compilation instructions for Debian/Ubuntu, Mac OS X and Windows, as well as precompiled packages. The source code is available at the official Github repository: <http://github.com/MTG/essentia>. In our current work we are focused on expanding the library and the community of users, and all active Essentia users are encouraged to contribute to the library.

## References

- [1] Serra, X., Magas, M., Benetos, E., Chudy, M., Dixon, S., Flexer, A., Gómez, E., Gouyon, F., Herrera, P., Jordà, S., Paytavi, O, Peeters, G., Schlüter, J., Vinet, H., and Widmer, G., Roadmap for Music Information ReSearch, G. Peeters, Ed., 2013. [Online].
- [2] Bogdanov, D., Wack N., Gómez E., Gulati S., Herrera P., Mayor O., Roma, G., Salamon, J., Zapata, J., Serra, X. (2013). ESSENTIA: an Audio Analysis Library for Music Information Retrieval. International Society for Music Information Retrieval Conference (ISMIR'13). 493-498.
- [3] Bogdanov, D., Wack N., Gómez E., Gulati S., Herrera P., Mayor O., Roma, G., Salamon, J., Zapata, J., Serra, X. (2013). ESSENTIA: an Open-Source Library for Sound and Music Analysis. ACM International Conference on Multimedia (MM'13).

# SIGMM Award for Outstanding PhD Thesis in Multimedia Computing, Communications and Applications

## Award Description

This award will be presented at most once per year to a researcher whose PhD thesis has the potential of very high impact in multimedia computing, communication and applications, or gives direct evidence of such impact. A selection committee will evaluate contributions towards advances in multimedia including multimedia processing, multimedia systems, multimedia network services, multimedia applications and interfaces. The award will recognize members of the SIGMM community and their research contributions in their PhD theses as well as the potential of impact of their PhD theses in multimedia area. The selection committee will focus on candidates' contributions as judged by innovative ideas and potential impact resulting from their PhD work.

The award includes a US\$500 honorarium, an award certificate of recognition, and an invitation for the recipient to receive the award at a current year's SIGMM-sponsored conference, the ACM International Conference on Multimedia (ACM Multimedia). A public



### 3. THE AGE OF SEMANTIC DESCRIPTORS

*There are four kinds of sounds of water: the sounds of cataracts, of gushing springs, of rapids, and of gulleys. There are three kinds of sounds of wind: the sounds of "pine waves," of autumn leaves, and of storm upon the water. There are two kinds of sounds of rain: the sounds of raindrops upon the leaves of wu'tung and lotus, and the sounds of rain water coming down from the eaves into bamboo pails.*

Lin Yutang, *The importance of living* (1937), p. 322.

*los animales se dividen en (a) pertenecientes al Emperador, (b) embalsamados, (c) amaestrados, (d) lechones, (e) sirenas, (f) fabulosos, (g) perros sueltos, (h) incluidos en esta clasificación, (i) que se agitan como locos, (j) innumerables, (k) dibujados con un pincel finísimo de pelo de camello, (l) etcétera, (m) que acaban de romper el jarrón, (n) que de lejos parecen moscas.*

Jorge Luis Borges, *El idioma analítico de John Wilkins*,  
Otras Inquisiciones (1937-1952)

#### 3.1. Introduction

The age of semantic descriptors started when researchers realized that low-level features, computed bottom-up (directly from the raw audio file or by means of some front-end, without any or very shallow domain knowledge included in the computation) do not yield the relationships, similarities, rankings or partitions that humans tend to perceive between objects. Similarity, one of the keystones in content analysis of multimedia, becomes brittle when only signal-based basic features are used to compute it. Moreover, humans tend to find similarities not only by differentially attending on specific perceptual dimensions or features but also by simplifying and compacting the surface features, by means of “categories” or “concepts” (Goldstone et al., 2017), and the concepts can be extremely diverse and inconsistent (Aucouturier & Pachet, 2002). Added to this picture is the fact that listeners tend to activate semantic associations when listening to music (Koelsch et al., 2004), so this will influence the type of concepts they use when describing music or searching for it. To deal with categories created and manipulated by humans implies addressing “meaning” or “semantics”. Semantic descriptors, also named “high-level descriptors” convey a defined meaning that is not necessarily directly associated with the signal properties, but with properties of subjective feelings and interpretations (from music theory, cognitive theory, or subjective/naïve theories about whatever knowledge domain, music in our case). Nack (2004) provides a complementary (though incomplete) summary of the problem: “*the major problem with this static approach to metadata [that is, the approach based on low-level features] is that it doesn’t reflect the continuous process of interpreting and understanding a media item’s syntactic and semantic concepts. Media items continue to be produced mainly for a particular purpose. A video sequence of a heart operation, for example, might be produced for an educational multimedia project. Yet, what hinders us from letting the same sequence feature*

*prominently in a soap opera, for example, where it might create a cliffhanger ending to an episode and inspire viewers to watch the next episode, too? The answer is nothing—aside from current technology, which prevents us from representing this change in a media item’s behavior. To support such flexibility schemata developers and users must agree upon semantic-based and machine-processable metadata collections during established media workflow practices.”*

Semantic descriptions must deal, then, with the so-called “semantic gap”, which is the cleft separating low-level features and the concepts humans use when dealing with audio and music (Celma et al., 2006). Machine learning provides many tools and techniques to bridge the semantic gap by means of mapping feature-based descriptions of music with categories or meanings given to them by humans. As some of these techniques call for sampling the ways humans categorize music, annotating collections of music is another activity that characterizes this age<sup>18</sup>. Annotated collections or “ground-truths” to be used to train machine learning models will pose many problems: the difficulty of doing open and replicable research when essential materials are subject to copyrights that prevent their redistribution or sharing, the number of annotations required (a time-consuming, error-prone task), and the reliability of such annotations (asking for several annotators working on the same material and checking for their consistency or inter-rater agreement). It is not surprising, then, to notice that some of the most used collections during this age contain questionable data (Sturm, 2012). If features (and not classification algorithms, as the myriad of papers “comparing classifiers” could make think to a naïve observer) are one of the building blocks of any serious research on MIR, the second building block is made of, for sure, well-annotated data. Several techniques have been proposed to ensure, increase or optimize data quality, or to reduce the required effort (Lessafre, 2006; Sordo et al., 2007; Humphrey et al., 2012; Peeters & Fort, 2012; Urbano & Schedl, 2013; Schlüter & Grill, 2015; McFee et al, 2015).

The list of semantic descriptors of music is long, though there are recurrent ones, coming from music theory and practice (tempo, meter, downbeat, chord, key, mode, swing, chorus, intro...) and from the emotional assessment of music (happy, sad, etc.). Lyrics would also be increasingly used since then (Logan et al., 2004; Laurier et al., 2008). Looking at the ways listeners tag music, their mental representations of music can be traced and mapped (if only in a rough and approximate way), and more “exotic” features can be hinted (some of them cannot be computed from the audio as “seen live”, “fun” or “bitch”, but many others like “warm”, “hollow”, “female vocal”, “metallic”, “melancholic”, “scenic”, and most of the genres and subgenre labels that are used all over the world, have acoustic correlates that can be formally computed and modelled). Papers selected for this chapter do not cover all the semantic fields addressed in our research, but I would like to leave a passing mention of at least an early work done on meter (i.e., double/triple meter detection) (Gouyon & Herrera, 2003), on tonality (Gómez & Herrera, 2004), on morphological descriptors (descriptors on the “shapes” and microfeatural changes of sounds, inspired by Pierre Schaeffer (1966)) (Ricard & Herrera, 2003; Ricard & Herrera, 2004; Cano et al., 2004a; Cano et al., 2004b)<sup>19</sup>, on subjective intensity categories (Sandvold & Herrera, 2005), on complexity (Streich & Herrera, 2006), danceability (Streich & Herrera, 2005), or on singing voice presence (Rocamora &

---

<sup>18</sup> see Murthy & Koolagudi (2018) for an almost exhaustive list of those collections, and see also <http://ismir.net/datasets.php>.

<sup>19</sup> See Peeters & Deruty (2010) for a thoroughful development of such descriptors.

Herrera, 2007)<sup>20</sup>. Some of the descriptors proposed in the previous papers were included in the Essentia library that was reported in the previous chapter.

Similarity has a special status as a kind of semantic representation in MIR. Long ago, Downie (2003) wrote: “the creation of rigorous and practicable theories concerning the nature of experiential similarity and relevance is the single most important challenge facing MIR researchers today” (p. 306). The assumption that we like, prefer, tend to, or know how to search for music using notions of similarity has frequently been taken for granted in many papers (Aucouturier & Pachet, 2002; Berenzweig, Logan, Ellis, & Whitman, 2004; Casey, 2002; Pampalk, 2006; Pohle et al., 2009). The assumption that, because two tracks or excerpts belong to the same artist or are linked by the same emotional category they must share similarity is another example of risky deduction. To conclude, the assumption that similarity happens out of any context may plague many similarity ratings taken at ground-truth value<sup>21</sup>

Similarity (though not “music similarity”) has also a special status and a long tradition of research in psychology and cognitive science. Because of that, it is surprising that during the age of semantic descriptors we were not able to take advantage of the techniques and models that those disciplines could provide, to advance in the understanding of music similarity, specially of audio music similarity. In the end we have, of course, developed techniques to model such similarity in a way that satisfies users’ needs on music search, listening and visualization (thanks to taking advantage of mining the way it is used in playlists, blogs, etc., i.e., thanks to considering “context”, as we will see in the next chapter) (McFee, 2012; Knees & Schedl, 2013; Schedl et al., 2014), and this functional apparent success has decreased its presence as a research topic. I am afraid though, that some basic research opportunities this topic could foster, taking advantage of the legacy we have in the experimental psychology literature (Goodman, 1972; Tversky, 1977; Garner, 1974; Kubovy, 1981; Smith, 1989; Medin et al., 1993; Goldstone, 1994, 1999; Fisher & Sloutsky, 2005), remain unexplored. The disappearing of the audio similarity task from recent MIREX editions should not be taken as a sign that the problem has been solved, but as an indicator that it is somehow stagnated.

## 3.2. Papers included in this chapter

Bogdanov, D., Serrà J., Wack N., **Herrera P.**, & Serra X. (2011). Unifying Low-level and High-level Music Similarity Measures. *IEEE Transactions on Multimedia*. 4, 687-701. (Journal h-index: 101; Journal IF 2016: 3.509; Q1 in Computer Science Applications journals; 72 citations)

Cano, P., Koppenberger, M., Le Groux, S., Ricard, J., Wack, N., **Herrera, P.** (2005). Nearest-neighbor sound annotation with a Wordnet taxonomy. *Journal of Intelligent Information Systems*, 24 (2), pp. 99-111. (Journal h-index: 47; Journal IF 2016; 1.107; Q2 in Information Systems journals; 20 citations)

---

<sup>20</sup> See Lee et al. (2018) for a recent review and state of the art approach.

<sup>21</sup> Fortunately, there is research cautiously enough on the pitfalls and requirements of research on music similarity (e.g., Chupchik, Rickert, & Mendelson, 1982; Lamont & Dibben, 2001; Pampalk, 2006; Jones et al., 2007; Urbano, 2013; Foster et al., 2014; Knees & Schedl, 2016).

Serrà, J., Gómez, E., **Herrera, P.**, Serra, X. (2008). Chroma binary similarity and local alignment applied to cover song identification. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(6), pp. 1138-1151. (Journal h-index: 91; Journal IF 2016: 2.491; Q1 in Acoustics and Ultrasonics journals; 245 citations)

Laurier, C., Meyers, O., Serrà, J., Blech, M., **Herrera, P.**, Serra, X. (2010). Indexing Music by Mood: Design and Integration of an Automatic Content-based Annotator. *Multimedia Tools and Applications*. 48(1), 161-184. (Journal h-index: 45; Journal IF 2016: 1.541; Q2 in Computer Networks and Communications journals; 40 citations)

Koelsch, S., Skouras S., Fritz T., **Herrera, P.**, Bonhage, C., Küssner, M. B. & Jacobs, A.M. (2013). The roles of superficial amygdala and auditory cortex in music-evoked fear and joy. *NeuroImage*. 81(1), 49-60. (Journal h-index: 307; Journal IF 2017; 5.426; Q1 in Cognitive Neuroscience journals; 79 citations)

### 3.3. Contributions in the selected papers

The age of semantic descriptors was framed, in my case, by an European project called SIMAC (Semantic Interaction with Music Audio Content)<sup>22</sup> that was an outstanding collaboration between Fabien Gouyon, Emilia Gómez, Xavier Serra and me in their conception, and then including the also outstanding partners from OFAI and QMUL (Herrera et al., 2005). This project was the first European project led by the MTG and I somehow played there the role of scientific director. Some papers included here or in the previous chapter were conceived or written out of that context or as refinements of that research effort. Of course, other research institutions and projects should be credited here as significantly pushing forward this quest for the semantics of music interaction, but this is left to proper historians of Science.

In “Unifying Low-level and High-level Music Similarity Measures” (Bogdanov et al., 2011) we presented improvements in our way to compute song similarity by means of obtaining a suitable distance measurement between songs represented on a rich featural space. In this paper, we proposed three of such distance measures based on the audio content: first, a low-level measure based on tempo-related description; second, a high-level semantic measure based on the inference of different musical dimensions by support vector machines. These dimensions include genre, culture, moods, instruments, rhythm, and tempo annotations. The third distance measure is a hybrid measure, which combines the above-mentioned distance measures with two existing low-level measures: a Euclidean distance based on principal component analysis of timbral, temporal, and tonal descriptors, and a timbral distance based on single Gaussian Mel-frequency cepstral coefficient (MFCC) modelling. We evaluated our proposed measures against several baseline measures. We did this objectively (using a comprehensive set of music collections) and subjectively (by means of listeners’ ratings). Results showed that the proposed methods achieved accuracies comparable to the baseline approaches in the case of the tempo and classifier-based measures. The highest accuracies, though, were obtained by the hybrid distance. Furthermore, the proposed classifier-based approach

---

<sup>22</sup> <http://mtg.upf.edu/simac/>, [https://cordis.europa.eu/project/rcn/71237\\_en.html](https://cordis.europa.eu/project/rcn/71237_en.html)

opened up the possibility to explore distance measures that are based on semantic notions and that would be further exploited in MIREX 2009<sup>23</sup>, where the hybrid algorithm scored the third best (though there were no statistically significant differences between the best three submissions) and in MIREX 2010 (Wack et al., 2010), again with results at the top of the list.

In “Nearest-neighbor sound annotation with a Wordnet taxonomy” (Cano et al., 2005) we presented a system allowing the semi-automatic annotation of sound effects. In this domain most of the annotation was done manually (a human editor created the appropriate taxonomy and assigned the tags to each file automatically, and therefore every sound content provider used a different taxonomy, with different tags). Automatic annotation can only work for reduced domains such as musical instruments, or car sounds, for example, as the number of potential categories to be labelled is so big (on the range of tens of thousands) that automatic classifiers could not cope with that, at the time of the reported research. We approached this scenario considering that a sound annotation/recognition tool would require, first, a taxonomy representing the complete sonic world and, then, a nearest-neighbour-based strategy for labelling sounds. To help with the annotation process (manual or automatic), we took advantage of Wordnet<sup>24</sup>, a semantic network or English lexical database, that organized real world knowledge (Miller, 1995). With Wordnet, relationships between concepts (and their tags) could be exploited to assign to each sound a rich semantic description, and to make possible efficient tag propagation between annotated and to-be-annotated sounds. In the reported evaluation a 30% of correct prediction was achieved on a database of over 50000 sounds, and more than 1600 concepts. It is worth to note that, in the times our paper was published, evaluations with such a big amount of sounds and categories were not frequent (or perhaps inexistent). Research here was motivated by our development of an industry-scale annotator to be used as the backbone for a huge collection of sound effects to be managed by “The Tape Gallery”<sup>25</sup>, at that time one of the leading companies in web-accessible sound effects. Wordnet was also used for the first time in MIR, as far as I know, and it has been further used for getting semantic similarity in different retrieval contexts (Davies & Plumbley, 2007; Mesaros et al., 2013, Mechtley, 2013). Another remarkable contribution is the knowledge we gained during the reported research, helping us to build search functions for Freesound<sup>26</sup>, the open database of sound recordings developed in the MTG.

One of the most difficult problems to be tackled in MIR is that of detecting versions or covers, a task whose popularity in the MIR community increased during the age of semantic descriptors, as it provides a direct and objective way to evaluate music similarity algorithms, and challenges our assumptions on what is the essence of music and what is an original work of art. The concept of cover is a sloppy one, as it can be a faithful copy almost identical to the original, or a musical creation that only preserves a hook, or a characteristic rhythm pattern. In fact, in the digital domain the concept of “original” virtually does not make sense (unless a trustable timestamp or watermark can be associated to it, copies cannot be distinguished from originals). The concept of invariance

---

<sup>23</sup> [http://www.music-ir.org/mirex/2009/index.php/Audio\\_Music\\_Similarity\\_and\\_Retrieval\\_Results](http://www.music-ir.org/mirex/2009/index.php/Audio_Music_Similarity_and_Retrieval_Results)

<sup>24</sup> <http://wordnet.princeton.edu>

<sup>25</sup> <https://www.mixonline.com/technology/tape-gallery-sound-effects-librarycom-371389>, note that the article is dated before our partnership with them.

<sup>26</sup> <http://www.freesound.org>

is very pertinent here as in many cases artists respect some invariant musical aspect (usually tonal, sometimes rhythmic, timbral, textual, etc.) in order their cover to qualify as such. Understanding the way listeners decide that something is a version of another musical excerpt, made possible to focus on a series of invariances (mostly tempo and tonality related) that some of our cover detection algorithms took advantage of. In “Chroma binary similarity and local alignment applied to cover song identification” (Serrà et al., 2008), we presented an innovative technique for audio signal comparison based on tonal subsequence alignment and we discussed and evaluated its application to detect cover versions (i.e., different performances of the same underlying musical piece), a topic that, as far as we know, we opened to the MIR community in 2006 (Gómez & Herrera, 2006). The selected paper, that two years ago became my most-cited one<sup>27</sup>, first presents a series of experiments carried out with two state-of-the-art methods for cover song identification. We studied several components of these (such as chroma resolution and similarity, transposition, beat tracking or Dynamic Time Warping constraints), to discover which characteristics would be desirable for a competitive cover song identifier. After analysing many cross-validated results, the importance of these characteristics was discussed, and the best-performing ones were finally applied to the newly proposed method. Multiple evaluations of this newly proposed method confirmed a large increase in identification accuracy when comparing it with alternative state-of-the-art approaches. The presented technique got the best results in the MIREX 2008 and 2009 campaigns, and it was later outperformed by another one which took advantage of complex networks techniques (Serrà et al., 2012).

A method for creating automatic music mood annotations was presented in “Indexing Music by Mood: Design and Integration of an Automatic Content-based Annotator” (Laurier et al., 2010). In addition to a complete evaluation of a music mood classifier that decided if the music transmitted happiness, sadness, anger or relax, we reported on the integration and subjective evaluation of a fast and scalable version of it in a large-scale annotator that was used as demonstrator of the European Project PHAROS<sup>28</sup>. The mood annotation method that we presented was inspired on results from psychological studies on emotion categorization and characterization, and was framed into a supervised learning approach using musical features automatically extracted from the raw audio signal. We discussed some of the most relevant audio features to solve this problem which, surprisingly, in some cases (anger, relax) were supported by timbre features and not by musically sophisticated concepts such as tonality. A ground truth, used for training, was created using both social network information systems (wisdom of crowds) and individual experts (wisdom of the few). At the experimental level, we evaluated our approach on a database of 1000 songs. Tests of different classification methods, configurations and optimizations were carried on, showing that Support Vector Machines provided the best classification model for the task at hand. Moreover, we evaluated the algorithm robustness against different audio compression schemes. This fact, often neglected, is fundamental to build a system that is usable in real conditions, as it was one of the requirements of the industrial partners of the project. In connection with this research, the Moodcloud demonstrator (Laurier & Herrera, 2008) was an original graphical tool showing how the emotions communicated by music changed as the music

---

<sup>27</sup> 248 citations, according to Google Scholar, retrieved on September, 21<sup>st</sup>, 2018.

<sup>28</sup> [https://cordis.europa.eu/result/rcn/193196\\_en.html](https://cordis.europa.eu/result/rcn/193196_en.html)

was being played. Mood classification models were used with very good results in MIREX 2007 (2<sup>nd</sup> out of 9 participants) and MIREX 2009 (4<sup>th</sup> out of more than 30 participants).

Semantic descriptors can be useful outside the typical scenarios in MIR (tag-based retrieval, similarity, recommendation, transcription...). Neuroscientists devoted to understanding our musical brain must use music stimuli that belong to certain categories, to study how the brain areas are activated by them, or how specific electrical currents change accordingly. While it can be quite unequivocal to decide if a music excerpt contains a vocal sound, or a major chord, or a deceptive cadence, it becomes trickier when “aggressive”, “sad” or “beautiful” music must be selected as stimulus. In that context, our contribution to “The roles of superficial amygdala and auditory cortex in music-evoked fear and joy” (Koelsch et al., 2013) was the validation, by means of the mood models reported in Laurier et al. (2010), of the used stimuli (i.e., to provide support on the sad or happy nature of the excerpts neuroscientists had selected as such), and the analysis of acoustic and musical features that correlated with such categories. The neuroscience findings of that paper are, of course, not relevant here. The paper though, motivates reflections on how the lack of normative collections or ground-truths affects not only MIR but other disciplines. It is surprising that music psychologists/neuroscientists have only recently created a “normative” collection of music (i.e., validated and carefully crafted to be taken as representative of certain features for a large population of subjects) to study music-related emotions (Imbir & Golab, 2016; but see also Anjalaki et al., 2017 for an alternative). A normative collection could be considered as a high-quality ground-truth collection, which has been validated with specific techniques and figures of merit that ensure it can be used as a reliable measurement device (of emotions communicated by music, for example). Because of the strict care and control put in its construction it is used routinely by many researchers when addressing the problem for which it is devised. Normative collections do not exist for frequently-used semantic descriptors, though they do for research on topics<sup>29</sup> such as objects in images (Brodeur et al., 2010) or emotional content of sounds (Bradley & Lang, 2007). Synergies between MIR researchers, psychologists and neuroscientists could make such normative collections possible. Unfortunately, this has never happened yet, to my knowledge.

Even though semantic descriptors made possible to detect and model essential concepts in music experiencing, bridging the semantic gap required something more than intensive statistical modelling or machine learning applied on feature-plenty descriptions of music items. As it happened in the age of feature extractors, shortcomings and glass-ceiling effects were evident as the techniques, user requirements and commercial successes pushed forward the available knowledge. The perspectives taken in the age of semantic descriptors overlooked the context where a person interacts with music contents, as the context may change the semantics involved. As a matter of example, music recommendation became a challenging topic that, only with the consideration of contextual factors could be tackled, if not effectively, at least promisingly. The age of context-aware systems was, then, ready to start.

---

<sup>29</sup> See, for example, [http://www.psychwiki.com/wiki/Archives\\_of\\_data\\_and\\_stimuli](http://www.psychwiki.com/wiki/Archives_of_data_and_stimuli), and <https://www.cogsci.nl/stimulus-sets>, although not many of the sets are truly normalized. For MIR purposes, very diverse sets can be found here: <https://www.audiocontentanalysis.org/data-sets/> and here: <https://www.ismir.net/resources.html>

Bogdanov, D., Serrà J., Wack N., **Herrera P.**, & Serra X. (2011). "Unifying Low-level and High-level Music Similarity Measures". IEEE Transactions on Multimedia. 4, 687-701.

DOI: [10.1109/TMM.2011.2125784](https://doi.org/10.1109/TMM.2011.2125784)

Print ISSN: 1520-9210

Online ISSN: 1941-0077



# Unifying Low-Level and High-Level Music Similarity Measures

Dmitry Bogdanov, Joan Serrà, Nicolas Wack, Perfecto Herrera, and Xavier Serra

**Abstract**—Measuring music similarity is essential for multimedia retrieval. For music items, this task can be regarded as obtaining a suitable distance measurement between songs defined on a certain feature space. In this paper, we propose three of such distance measures based on the audio content: first, a low-level measure based on tempo-related description; second, a high-level semantic measure based on the inference of different musical dimensions by support vector machines. These dimensions include genre, culture, moods, instruments, rhythm, and tempo annotations. Third, a hybrid measure which combines the above-mentioned distance measures with two existing low-level measures: a Euclidean distance based on principal component analysis of timbral, temporal, and tonal descriptors, and a timbral distance based on single Gaussian Mel-frequency cepstral coefficient (MFCC) modeling. We evaluate our proposed measures against a number of baseline measures. We do this objectively based on a comprehensive set of music collections, and subjectively based on listeners' ratings. Results show that the proposed methods achieve accuracies comparable to the baseline approaches in the case of the tempo and classifier-based measures. The highest accuracies are obtained by the hybrid distance. Furthermore, the proposed classifier-based approach opens up the possibility to explore distance measures that are based on semantic notions.

**Index Terms**—Distance measurement, information retrieval, knowledge acquisition, multimedia computing, multimedia databases, music.

## I. INTRODUCTION

RAPID development of digital technologies, the Internet, and the multimedia industry have provoked a huge excess of information. An increasingly growing amount of multimedia data complicates search, retrieval, and recommendation of relevant information. For example, in the digital music industry, major Internet stores such as the iTunes Store contain up to 14 million songs,<sup>1</sup> adding thousands of new songs

every month. In such circumstances, fast and efficient retrieval approaches operating on large-scale multimedia databases are necessary [1]. Specifically, similarity search is a challenging scientific problem, which helps to facilitate advances in multimedia knowledge, organization, and recommendation. Therefore, it can serve the user's needs and satisfaction within educative, explorative, social, and entertainment multimedia applications.

Studying the ways to search and recommend music to a user is a central task within the music information retrieval (MIR) community [2]. From a simplistic point of view, this task can be regarded as obtaining a suitable distance<sup>2</sup> measurement between a query song and a set of potential candidates. This way, one maps these songs to a certain feature space where a dissimilarity measure can be computed. Currently, researchers and practitioners fill in this feature space with information extracted from the audio content,<sup>3</sup> context, or both. Contextual information, in the form of user ratings [3] and social tags [4], is a powerful source for measuring music similarity. However, it becomes problematic to obtain such data in a long-tail [5]. General lack of user ratings and social tags for unpopular multimedia items complicate their sufficient characterization, as multimedia consumption is biased towards popular items. Alternatively, information extracted from the audio content can help to overcome this problem [6].

The present work deals with content-based approaches to music similarity. We organize this paper into three parts, dealing with the state-of-the-art, the proposal of two simple distance measurements, and the proposal of a hybrid (non-simple) distance measurement, respectively.

In the first part (Section II), we review related state-of-the-art, including current approaches to music similarity (Section II-A) and low-level audio descriptors available to our research (Section II-B). Furthermore, we briefly explain a number of existing simple approaches, which we use as a baseline for evaluating our proposed methods. Throughout the paper, we assume simple approaches to be those which are not constituted by a number of distances.<sup>4</sup> More concretely, as baseline approaches, we consider Euclidean distances defined on sets of timbral, rhythmic, and tonal descriptors (Sections II-CI and Section II-C-I) and Kullback-Leibler divergence defined on Gaussian mixture models (GMMs) of Mel-frequency cepstral coefficients (MFCCs, Section II-CIII).

<sup>2</sup>We here pragmatically use the term "distance" to refer to any dissimilarity measurement between songs.

<sup>3</sup>We pragmatically use the term "content" to refer to any information extracted from the audio signal.

<sup>4</sup>We have opted for the term "simple" instead of other appropriate terms, such as "non-hybrid" and "homogeneous".

Manuscript received November 15, 2010; revised February 17, 2011; accepted February 23, 2011. Date of publication March 10, 2011; date of current version July 20, 2011. This work was supported in part by the FI Grant of Generalitat de Catalunya (AGAUR); in part by the Music 3.0 project of the Spanish Ministry of Industry, Tourism, and Trade (Avanza Contenidos, TSI-070100-2008-318); and in part by the Buscamedia project (CEN-20091026). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Shrikanth Narayanan.

The authors are with the Music Technology Group, Universitat Pompeu Fabra, 08018 Barcelona, Spain (e-mail: dmitry.bogdanov@upf.edu; joan.serraj@upf.edu; nicolas.wack@upf.edu; perfecto.herrera@upf.edu; xavier.serra@upf.edu).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TMM.2011.2125784

<sup>1</sup>[http://en.wikipedia.org/wiki/iTunes\\_Store](http://en.wikipedia.org/wiki/iTunes_Store)

In the second part, which we partially presented in [7], we compare the aforementioned baseline approaches against two novel distance measures (Section III). The first idea we explore consists of the use of tempo-related musical aspects. We propose a distance based on two low-level rhythmic descriptors, namely beats per minute and onset rate (Section III-A). The second idea we explore shifts the problem to a more high-level (semantic) domain as we propose to use high-level semantic dimensions, including information about genre and musical culture, moods and instruments, and rhythm and tempo. With regard to this aspect, we continue the research of [8]–[10] but, more in the line of [10], we investigate the possibility of benefiting from results obtained in different classification tasks and transferring this acquired knowledge to the context of music similarity (Section III-B). More specifically, as our first main technical contribution, we infer different groups of musical dimensions by using support vector machines and use a high-level modular distance which combines these dimensions. Among the qualities of this classifier-based distance, we strive for high modularity, being able to easily append additional dimensions. Moreover, we strive for descriptiveness, being able to explain similarity to a user.

We evaluate all the considered simple approaches with a uniform methodological basis, including an objective evaluation on several comprehensive ground truth music collections (Section IV-A) and a subjective evaluation based on ratings given by real listeners (Section IV-C). We show that, in spite of being conceptually different, the proposed methods achieve comparable or even higher accuracies than the considered baseline approaches (Sections IV-B and Section IV-D). Finally, we illustrate the benefits of the proposed classifier-based distance for music similarity justification to a user (Section V). In addition, we demonstrate an example of possible semantic explanation of similarity between songs.

In the third part, we explore the possibility of creating a hybrid approach, based on the considered simple approaches as potential components. As our second main technical contribution, we propose a new distance measure that combines a low-level Euclidean distance based on principal component analysis (PCA), a timbral distance based on single Gaussian MFCC modeling, and our proposed tempo-based and semantic classifier-based distances (Section VI). These choices are motivated by the results obtained in the subjective evaluation of simple approaches performed in the second part of the paper. We hypothesize that such combination of conceptually different approaches, covering timbral, rhythmic, and semantic aspects of music similarity, is more appropriate from the point of view of music cognition [11] and, thus, it could lead to a better performance from the point of view of the listener. Indeed, a number of works support this idea though being limited by combining only timbral and rhythmic aspects into a hybrid distance [12]–[17], and, alternatively, timbral and tonal, or timbral and semantic ones [18]. To the best of the authors' knowledge, more extended combinations of timbral, rhythmic, tonal, and semantic dimensions, providing a single hybrid distance, have not yet been studied.

We evaluate the hybrid approach against its component approaches objectively, performing a cross-collection out-of-sample test on two large-scale music collections

(Section VII-A), and subjectively, based on ratings of 21 real listeners (Section VII-C). We find that the proposed hybrid method reaches a better performance than all considered approaches, both objectively (Section VII-B) and subjectively (Section VII-D). We subjectively evaluate our classifier-based and hybrid approaches against a number of state-of-the-art distance measures within the bounds of an international evaluation framework (Section VIII-A). Notably, our hybrid approach is found to be one of the best performing participants (Section VIII-B). We finally state general conclusions and discuss the possibility of further improvements (Section IX).

## II. SCIENTIFIC BACKGROUND

### A. Music Similarity

Focusing on audio content-based similarity, there exist a wide variety of approaches for providing a distance measurement between songs. These approaches comprise both the selection of audio descriptors and the choice of an appropriate distance function. Representing the songs as points in a feature space with an  $L_p$  metric is a straightforward approach. Cano *et al.* [19] demonstrate such an approach using a Euclidean metric after a PCA transformation of a preliminary selected combination of timbral, temporal, and tonal descriptors. Similarly, Slaney *et al.* [20] apply a Euclidean metric on loudness and temporal descriptors, and use a number of algorithms to improve performance. These algorithms include whitening transformation, linear discriminant analysis (LDA), relevant component analysis (RCA) [21], neighborhood components analysis, and large-margin nearest neighbor classification [22].

As well, specific timbral representations exist, the most prominent one being modeling the songs as clouds of vectors of MFCCs, calculated on a frame basis. Logan and Salomon [23] represent such clouds as cluster models, comparing them with the Earth mover's distance. Mandel and Ellis [24] compare means and covariances of MFCCs applying the Mahalanobis distance. Furthermore, GMMs can be used to represent the clouds as probability distributions, and then these distributions can be compared by the symmetrized Kullback-Leibler divergence. However, in practice, approximations are required for the case of several Gaussian components in a mixture. To this end, Aucouturier *et al.* [25], [26] compare GMMs by means of Monte Carlo sampling. In contrast, Mandel and Ellis [24] and Flexer *et al.* [27] simplify the models to single Gaussian representations, for which a closed form of the Kullback-Leibler divergence exists. Pampalk [13] gives a global overview of these approaches. As well, Jensen *et al.* [28] provide an evaluation of different GMM configurations. Besides MFCCs, more descriptors can be used for timbral distance measurement. For example, Li and Ogihara [29] apply a Euclidean metric on a set of descriptors, including Daubechies wavelet coefficient histograms.

Temporal (or rhythmic) representation of the songs is another important aspect. A number of works propose specific temporal distances in combination with timbral ones. For example, Pampalk *et al.* [12], [13] exploit fluctuation patterns, which describe spectral fluctuations over time, together with several derivative

descriptors, modeling overall tempo and fluctuation information at specific frequencies. A hybrid distance is then defined as a linear combination of a Euclidean distance on fluctuation patterns together with a timbral distance, based on GMMs of MFCCs. Pohle *et al.* [14] follow this idea, but propose using a cosine similarity distance for fluctuation patterns together with a specific distance measure related to cosine similarity for GMMs of MFCCs. Furthermore, they propose an alternative temporal descriptor set, including a modification of fluctuation patterns (onset patterns and onset coefficients), and additional timbral descriptors (spectral contrast coefficients, harmoniousness, and attackness) along with MFCCs for single Gaussian modeling [15], [16]. Song and Zhang [17] present a hybrid distance measure, combining a timbral Earth mover's distance on MFCC cluster models, a timbral Euclidean distance on spectrum histograms, and a temporal Euclidean distance on fluctuation patterns.

Finally, some attempts to exploit tonal representation of songs exist. Ellis and Poliner [30], Marolt [31], and Serrà *et al.* [32] present specific melodic and tonality distance measurements, not addressed to the task of music similarity, but to version (cover) identification. In principle, their approaches are based on matching sequences of pitch class profiles, or chroma feature vectors, representing the pitch class distributions (including the melody) for different songs.

Though common approaches for content-based music similarity may include a variety of perceptually relevant descriptors related to different musical aspects, such descriptors are, in general, relatively low-level and not directly associated with a semantic explanation [33]. In contrast, research on computing high-level semantic features from low-level audio descriptors exists. In particular, in the context of MIR classification problems, genre classification [34], mood detection [35], [36], and artist identification [24] have gathered much research attention.

Starting from the relative success of this research, we hypothesize that the combination of classification problem outputs can be a relevant step to overcome the so-called semantic gap [33] between human judgements and low-level machine learning inferences, specifically in the case of content-based music similarity. A number of works support this hypothesis. Berenzweig *et al.* [9] propose to infer high-level semantic dimensions, such as genres and "canonical" artists, from low-level timbral descriptors, such as MFCCs, by means of neural networks. The inference is done on a frame basis, and the resulting clouds in high-level feature space are compared by centroids with a Euclidean distance. Barrington *et al.* [8] train GMMs of MFCCs for a number of semantic concepts, such as genres, moods, instrumentation, vocals, and rhythm. Thereafter, high-level descriptors can be obtained by computing the probabilities of each concept on a frame basis. The resulting semantic clouds of songs can be represented by GMMs, and compared with Kullback-Leibler divergence. McFree and Lanckriet [18] propose a hybrid low-dimensional feature transformation embedding musical artists into Euclidean space subject to a partial order, based on a set of manually annotated artist similarity triplets, over pairwise low-level and semantic distances. As such, the authors consider low-level timbral distance, based on MFCCs, tonal distance, based on chroma descriptors, and the above-mentioned semantic distance [8]. The evaluation includes the embeddings,

which merge timbral and tonal distances, and, alternatively, timbral and semantic distances. West and Lamere [10] apply classifiers to infer semantic features of the songs. In their experiment, Mel-frequency spectral irregularities are used as an input for a genre classifier. The output class probabilities form a new high-level feature space, and are compared with a Euclidean distance. The authors propose to use classification and regression trees or LDA for classification.

In spite of having a variety of potential content-based approaches to music similarity, still there exist certain open issues. The distances, operating solely on low-level audio descriptors, lack semantic explanation of similarity on a level which human judgements operate. The majority of approaches, both low-level and high-level, focus mostly on timbral descriptors, whereas other types of low-level descriptors, such as temporal and tonal, are potentially useful as well. Furthermore, comparative evaluations are necessary, especially those carried out comprehensively and uniformly on large music collections. In existing research, there is a lack of such comparative evaluations, taking into consideration different approaches. Objective evaluation criteria of music similarity are generally reduced to co-occurrences of genre, album, and artist labels, being tested on relatively small ground truth collections. In turn, subjective evaluations with human raters are not common. We will focus on filling in these open issues, employing comprehensive music collections, objective criteria for similarity, and human listeners for subjective evaluations. As existing approaches still perform relatively poorly, we hypothesize that better performance may be achieved by combining conceptually different distance measurements, which will help to jointly exploit different aspects of music similarity.

## B. Musical Descriptors

In the present work, we characterize each song using an in-house audio analysis tool.<sup>5</sup> From this tool, we use 59 descriptor classes in total, characterizing global properties of songs, and covering timbral, temporal, and tonal aspects of musical audio. The majority of these descriptors are extracted on a frame-by-frame basis with a 46 ms frame size, and 23 ms hop size, and then summarized by their means and variances across these frames. In the case of multidimensional descriptors, covariances between components are also considered (e.g., with MFCCs). Since it is not the objective of this paper to review existing methods for descriptor extraction, we just provide a brief overview of the classes we use in Table I. The interested reader is referred to the cited literature for further details.

## C. Baseline Simple Approaches

In this work, we consider a number of conceptually different simple approaches to music similarity. Among them we indicate several baselines, which will be used in objective and subjective evaluations, and moreover will be regarded as potential components of the hybrid approach.

1) *Euclidean Distance Based on Principal Component Analysis  $L_2$ -PCA*: As a starting point, we follow the ideas proposed by Cano *et al.* [19], and apply an unweighted Euclidean metric

<sup>5</sup><http://mtg.upf.edu/technologies/essentia>

TABLE I  
OVERVIEW OF MUSICAL DESCRIPTORS

Descriptor group	Descriptor class
Timbral	Bark bands [35], [37]
	MFCCs [13], [35], [37], [38]
	Pitch [39], pitch centroid [40]
	Spectral centroid, spread, kurtosis, rolloff, decrease, skewness [35], [37], [41]
	High-frequency content [39], [41]
	Spectral complexity [35]
	Spectral crest, flatness, flux [37], [41]
	Spectral energy, energy bands, strong peak, tristimulus [41]
Rhythmic	Inharmonicity, odd to even harmonic energy ratio [37]
	BPM, onset rate [35], [39], [41]
Tonal	Beats loudness, beats loudness bass [40]
	Transposed and untransposed harmonic pitch class profiles, key strength [35], [42]
	Tuning frequency [42]
	Dissonance [35], [43]
	Chord change rate [35]
Miscellaneous	Chords histogram, equal tempered deviations, non-tempered/tempered energy ratio, diatonic strength [40]
	Average loudness [37]
	Zero-crossing rate [13], [37]

on a manually selected subset of the descriptors outlined above,<sup>6</sup> This subset includes bark bands, pitch, spectral centroid, spread, kurtosis, rolloff, decrease, skewness, high-frequency content, spectral complexity, spectral crest, flatness, flux, spectral energy, energy bands, strong peak, tristimulus, inharmonicity, odd to even harmonic energy ratio, beats loudness, beats loudness bass, untransposed harmonic pitch class profiles, key strength, average loudness, and zero-crossing rate.

Preliminary steps include descriptor normalization in the interval  $[0, 1]$  and PCA [44] to reduce the dimension of the descriptor space to 25 variables. The choice of the number of target variables is conditioned by a trade-off between target descriptiveness and the curse of high-dimensionality [45]–[47], typical for  $L_p$  metrics, and is supported by research work on dimension reduction for music similarity [48]. Nevertheless, through our PCA dimensionality reduction, an average of 78% of the information variance was preserved on our music collections, reducing the number of 201 native descriptors by a factor of 8.

2) *Euclidean Distance Based on Relevant Component Analysis ( $L_2$ -RCA-1 and  $L_2$ -RCA-2)*: Along with the previous measure, we consider more possibilities of descriptor selection. In particular, we perform relevant component analysis (RCA) [21]. Similar to PCA, RCA gives a rescaling linear transformation of a descriptor space but is based on preliminary training on a number of groups of similar songs. Having such training data, the transformation reduces irrelevant variability in the data while amplifying relevant variability. As in the  $L_2$ -PCA approach, the output dimensionality is chosen to be 25. We consider both the descriptor subset used in  $L_2$ -PCA and the full descriptor set of Table I ( $L_2$ -RCA-1 and  $L_2$ -RCA-2, respectively).

3) *Kullback-Leibler Divergence Based on GMM of MFCCs (1G-MFCC)*: Alternatively, we consider timbre modeling with GMM as another baseline approach [26]. We implement the

<sup>6</sup>Specific details not included in the cited reference were consulted with P. Cano in personal communication.

simplification of this timbre model using single Gaussian with full covariance matrix [24], [27], [49]. Comparative research of timbre distance measures using GMMs indicates that such simplification can be used without significantly decreasing performance while being computationally less complex [13], [28]. As a distance measure between single Gaussian models for songs  $X$  and  $Y$ , we use a closed form symmetric approximation of the Kullback-Leibler divergence

$$\begin{aligned}
 d(X, Y) = & Tr(\Sigma_X^{-1}\Sigma_Y) + Tr(\Sigma_Y^{-1}\Sigma_X) \\
 & + Tr((\Sigma_X^{-1} + \Sigma_Y^{-1}) \\
 & \quad (\mu_X - \mu_Y)(\mu_X - \mu_Y)^T) \\
 & - 2N_{MFCC}
 \end{aligned} \tag{1}$$

where  $\mu_X$  and  $\mu_Y$  are MFCC means,  $\Sigma_X$  and  $\Sigma_Y$  are MFCC covariance matrices, and  $N_{MFCC}$  is the dimensionality of the MFCCs. This dimensionality can vary from 10 to 20 [28], [35], [50]. To preserve robustness against different audio encodings, the first 13 MFCC coefficients are taken [51].

### III. PROPOSED SIMPLE APPROACHES

Concerning simple approaches to music similarity, here we propose two novel distance measures that are conceptually different than what has been reviewed. We regard both approaches as potential components of the hybrid approach.

#### A. Tempo-Based Distance (TEMPO)

The first approach we propose is related to the exploitation of tempo-related musical aspects with a simple distance measure. This measure is based on two descriptors, beats per minute (BPM) and onset rate (OR), the latter representing the number of onsets per second. These descriptors are fundamental for the temporal description of music. Among different implementations, we opted for BPM and OR estimation algorithms presented in [39].

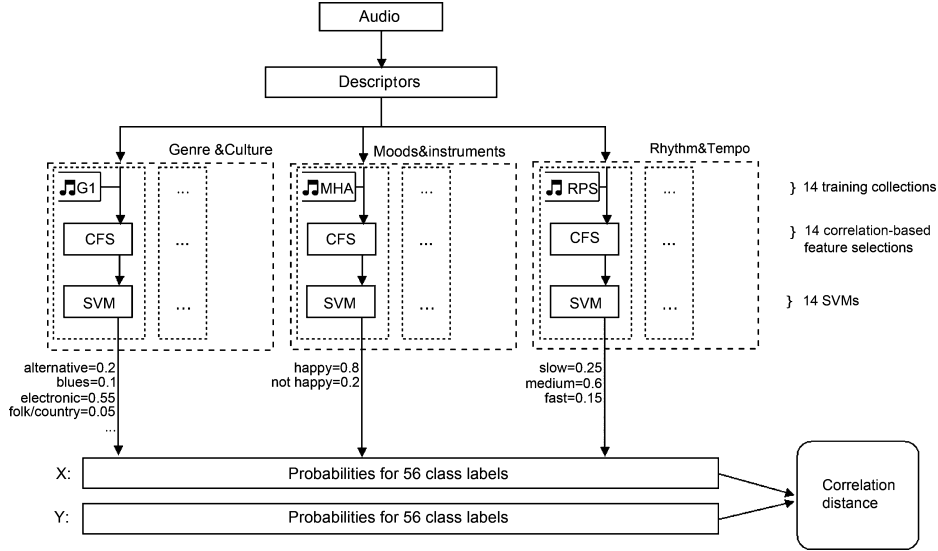


Fig. 1. General schema of CLAS distance. Given two songs  $X$  and  $Y$ , low-level audio descriptors are extracted, a number of SVM classifications are run based on ground truth music collections, and high-level representations, containing probabilities of classes for each classifier, are obtained. A distance between  $X$  and  $Y$  is calculated with correlation distances such as Pearson correlation distance.

For two songs  $X$  and  $Y$  with BPMs  $X_{\text{BPM}}$  and  $Y_{\text{BPM}}$ , and ORs  $X_{\text{OR}}$  and  $Y_{\text{OR}}$ , respectively, we determine a distance measure by a linear combination of two separate distance functions

$$d(X, Y) = w_{\text{BPM}}d_{\text{BPM}}(X, Y) + w_{\text{OR}}d_{\text{OR}}(X, Y) \quad (2)$$

defined for BPM as

$$d_{\text{BPM}}(X, Y) = \min_{i \in \mathbb{N}} \left( \alpha_{\text{BPM}}^{i-1} \left| \frac{\max(X_{\text{BPM}}, Y_{\text{BPM}})}{\min(X_{\text{BPM}}, Y_{\text{BPM}})} - i \right| \right) \quad (3)$$

and for OR as

$$d_{\text{OR}}(X, Y) = \min_{i \in \mathbb{N}} \left( \alpha_{\text{OR}}^{i-1} \left| \frac{\max(X_{\text{OR}}, Y_{\text{OR}})}{\min(X_{\text{OR}}, Y_{\text{OR}})} - i \right| \right) \quad (4)$$

where  $X_{\text{BPM}}, Y_{\text{BPM}}, X_{\text{OR}}, Y_{\text{OR}} > 0$ ,  $\alpha_{\text{BPM}}, \alpha_{\text{OR}} \geq 1$ . The parameters  $w_{\text{BPM}}$  and  $w_{\text{OR}}$  of (2) define the weights for each distance component. Equations (3) and (4) are based on the assumption that songs with the same BPMs (ORs) or multiples of the BPM (OR), e.g.,  $X_{\text{BPM}} = iY_{\text{BPM}}$ , are more similar than songs with non-multiple BPMs (ORs). For example, the songs  $X$  and  $Y$  with  $X_{\text{BPM}} = 140$  and  $Y_{\text{BPM}} = 70$  should have a closer distance than the songs  $X$  and  $Z$  with  $Z_{\text{BPM}} = 100$ . Our assumption is motivated by research on the perceptual effects of double or half tempo [52]. The strength of this assumption depends on the parameter  $\alpha_{\text{BPM}}(\alpha_{\text{OR}})$ . Moreover, such a distance can be helpful in relation to the common problem of tempo duplication (or halving) in automated tempo estimation [53], [54]. In the case of  $\alpha_{\text{BPM}} = 1$ , all multiple BPMs are treated equally, while in the case of  $\alpha_{\text{BPM}} > 1$ , preference inversely decreases with  $i$ . In practice, we use  $i = 1, 2, 4, 6$ .

Equations (2)–(4) formulate the proposed distance in the general case. In a parameter-tuning phase, we performed a

grid search with one of the ground truth music collections (RBL) under the objective evaluation criterion described in Section IV-A. Using this collection, which is focused on rhythmic aspects and contains songs with various rhythmic patterns, we found  $w_{\text{BPM}} = w_{\text{OR}} = 0.5$  and  $\alpha_{\text{BPM}} = \alpha_{\text{OR}} = 30$  to be the most plausible parameter configuration. Such values reveal the fact that in reality, both components are equally meaningful and that mainly a one-to-one relation of BPMs (ORs) is relevant for the music collection and descriptors we used to evaluate such rhythmic similarity. When our BPM (OR) estimator has increased duplicity errors (e.g., a BPM of 80 was estimated as 160), we should expect lower  $\alpha$  values.

#### B. Classifier-Based Distance (CLAS)

The second approach we propose derives a distance measure from diverse classification tasks. In contrast to the aforementioned methods, which directly operate on a low-level descriptor space, we first infer high-level semantic descriptors using suitably trained classifiers and then define a distance measure operating on this newly formed high-level semantic space. A schema of the approach is presented in Fig. 1.

For the first step, we choose standard multi-class support vector machines (SVMs) [44], which are shown to be an effective tool for different classification tasks in MIR [24], [35], [36], [55], [56]. We apply SVMs to infer different groups of musical dimensions such as 1) genre and musical culture, 2) moods and instruments, and 3) rhythm and tempo. To this end, 14 classification tasks are run according to all available ground truth collections presented in Table II. More concretely, we train one SVM per each ground truth collection, providing its annotated songs as a training input. For each collection and the corresponding SVM, a preliminary correlation-based feature

TABLE II  
GROUND TRUTH MUSIC COLLECTIONS EMPLOYED FOR OBJECTIVE EVALUATION OF THE SIMPLE APPROACHES. ALL PRESENTED COLLECTIONS ARE USED FOR TRAINING CLAS-BASED DISTANCES, EXCEPT G3, ART, AND ALB COLLECTIONS DUE TO INSUFFICIENT SIZE OF THEIR CLASS SAMPLES

Acronym	Category	Classes (musical dimensions)	Size	Source
G1	Genre & Culture	Alternative, blues, electronic, folk/country, funk/soul/rnb, jazz, pop, rap/hiphop, rock	1820 song excerpts, 46 - 490 per genre	[62]
G2	Genre & Culture	Classical, dance, hip-hop, jazz, pop, rhythm'n'blues, rock, speech	400 full songs, 50 per genre	In-house
G3	Genre & Culture	Alternative, blues, classical, country, electronica, folk, funk, heavy metal, hip-hop, jazz, pop, religious, rock, soul	140 full songs, 10 per genre	[63]
G4	Genre & Culture	Blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, rock	993 song excerpts, 100 per genre	[34]
CUL	Genre & Culture	Western, non-western	1640 song excerpts, 1132/508 per class	[55]
MHA	Moods & Instruments	Happy, non-happy	302 full songs + excerpts, 139/163 per class	[36] + in-house
MSA	Moods & Instruments	Sad, non-sad	230 full songs + excerpts, 96/134 per class	[36] + in-house
MAG	Moods & Instruments	Aggressive, non-aggressive	280 full songs + excerpts, 133/147 per class	[36] + in-house
MRE	Moods & Instruments	Relaxed, non-relaxed	446 full songs + excerpts, 145/301 per class	[36] + in-house
MPA	Moods & Instruments	Party, non-party	349 full songs + excerpts, 198/151 per class	In-house
MAC	Moods & Instruments	Acoustic, non-acoustic	321 full songs + excerpts, 193/128 per class	[36] + in-house
MEL	Moods & Instruments	Electronic, non-electronic	332 full songs + excerpts, 164/168 per class	[36] + in-house
MVI	Moods & Instruments	Voice, instrumental	1000 song excerpts, 500 per class	In-house
ART	Artist	200 different artist names	2000 song excerpts, 10 per artist	In-house
ALB	Album	200 different album titles	2000 song excerpts, 10 per album	In-house
RPS	Rhythm & Tempo	Perceptual speed: slow, medium, fast	3000 full songs, 1000 per class	In-house
RBL	Rhythm & Tempo	Chachacha, jive, quickstep, rumba, samba, tango, viennese waltz, waltz	683 song excerpts, 60 - 110 per class	[64]

selection (CFS) [44] over all available  $[0, 1]$ -normalized descriptors (Section II-B) is performed to optimize the descriptor selection for this particular classification task. As an output, the classifier provides probability values of classes on which it was trained. For example, a classifier using the G1 collection is trained on an optimized descriptor space, according to the collection's classes and the CFS process, and returns genre probabilities for the labels "alternative", "blues", "electronic", "folk/country", etc. Altogether, the classification results form a high-level descriptor space, which contains the probability values of each class for each SVM. Based on results in [35], we decided to use the libSVM<sup>7</sup> implementation with the C-SVC method and a radial basis function kernel with default parameters.

For the second step, namely defining a distance operating on a formed high-level semantic space (i.e., the one of the label probabilities), we consider different measures frequently used in collaborative filtering systems. Among the standard ones, we select the cosine distance (CLAS-Cos), Pearson correlation distance (CLAS-Pears) [5], [57], and Spearman's rho correlation distance (CLAS-Spear) [58]. Moreover, we consider a number of more sophisticated measures. In particular, the adjusted cosine distance (CLAS-Cos-A) [5], [57] is computed by taking into account the average probability for each class, i.e., compensating distinction between classifiers with different numbers of classes. Weighted cosine distance (CLAS-Cos-W) [59] and weighted Pearson correlation distance (CLAS-Pears-W) [60] are both weighted manually ( $W_M$ ) and also based on classification accuracy ( $W_A$ ). For  $W_M$ , we split the collections into three groups of musical dimensions, namely genre and musical

culture, moods and instruments, and rhythm and tempo. We empirically assign weights 0.50, 0.30, and 0.20, respectively. Our choice is supported by research on the effect of genre in terms of music perception [11], [61] and the fact that genre is the most common aspect of similarity used to evaluate distance measures in the MIR community [12]. For  $W_A$ , we evaluate the accuracy of each classifier, and thereafter assign proportional weights which sum to 1.

With this setup, the problem of content-based music similarity can be seen as a collaborative filtering problem of item-to-item similarity [57]. Such a problem can generally be solved by calculating a correlation distance between rows of a song/user rating matrix with the underlying idea that similar items should have similar ratings by certain users. Transferring this idea to our context, we can state that similar songs should have similar probabilities of certain classifier labels. To this extent, we compute song similarity on a song/user rating matrix with class labels playing the role of users, and probabilities playing the role of user ratings, so that each  $N$ -class classifier corresponds to  $N$  users.

#### IV. EVALUATION OF SIMPLE APPROACHES

We evaluated all considered approaches with a uniform methodological basis, including an objective evaluation on comprehensive ground truths and a subjective evaluation based on ratings given by real listeners. As an initial benchmark for the comparison of the considered approaches, we used a random distance (RAND), i.e., we selected a random number from the standard uniform distribution as the distance between two songs.

<sup>7</sup><http://www.csie.ntu.edu.tw/~cjlin/libsvm/>

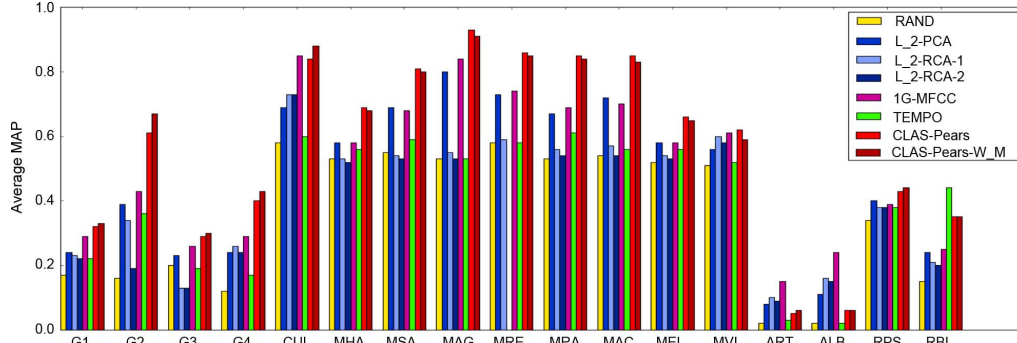


Fig. 2. Objective evaluation results (MAP) of the simple approaches for the different music collections considered.

TABLE III  
OBJECTIVE EVALUATION RESULTS (MAP) OF THE SIMPLE APPROACHES FOR THE DIFFERENT MUSIC COLLECTIONS CONSIDERED. N.C. STANDS FOR “NOT COMPUTED” DUE TO TECHNICAL DIFFICULTIES. FOR EACH COLLECTION, THE MAPS OF THE APPROACHES, WHICH PERFORM BEST WITHOUT STATISTICALLY SIGNIFICANT DIFFERENCE BETWEEN THEM, ARE MARKED IN BOLD

Method	G1	G2	G3	G4	CUL	MHA	MSA	MAG	MRE	MPA	MAC	MEL	MVI	ART	ALB	RPS	RBL
RAND	0.17	0.16	0.20	0.12	0.58	0.53	0.55	0.53	0.58	0.53	0.54	0.52	0.51	0.02	0.02	0.34	0.15
$L_2$ -PCA	0.24	0.39	0.23	0.24	0.69	0.58	0.69	0.80	0.73	0.67	0.72	0.58	0.56	0.08	0.11	0.40	0.24
$L_2$ -RCA-1	0.23	0.34	0.13	0.26	0.73	0.53	0.54	0.55	0.59	0.56	0.57	0.54	0.60	0.10	0.16	0.38	0.21
$L_2$ -RCA-2	0.22	0.19	0.13	0.24	0.73	0.52	0.53	0.53	N.C.	0.54	0.54	0.53	0.58	0.09	0.15	0.38	0.20
1G-MFCC	0.29	0.43	0.26	0.29	0.85	0.58	0.68	0.84	0.74	0.69	0.70	0.58	0.61	<b>0.15</b>	<b>0.24</b>	0.39	0.25
TEMPO	0.22	0.36	0.19	0.17	0.60	0.56	0.59	0.53	0.58	0.61	0.56	0.56	0.52	0.03	0.02	0.38	<b>0.44</b>
CLAS-Pears	0.32	0.61	0.29	0.40	0.84	<b>0.69</b>	<b>0.81</b>	<b>0.93</b>	<b>0.86</b>	<b>0.85</b>	<b>0.85</b>	<b>0.66</b>	<b>0.62</b>	0.05	0.06	0.43	0.35
CLAS-Pears- $W_M$	<b>0.33</b>	<b>0.67</b>	<b>0.30</b>	<b>0.43</b>	<b>0.88</b>	<b>0.68</b>	0.80	0.91	<b>0.85</b>	<b>0.84</b>	<b>0.83</b>	<b>0.65</b>	0.59	0.06	0.06	<b>0.44</b>	0.35

### A. Objective Evaluation Methodology

In our evaluations, we covered different musical dimensions such as genre, mood, artist, album, culture, rhythm, or presence or absence of voice. A number of ground truth music collections (including both full songs and excerpts) were employed for that purpose, and are presented in Table II. For some dimensions we used existing collections in the MIR field [34], [36], [55], [62]–[64], while for other dimensions we created manually labeled in-house collections. For each collection, we considered songs from the same class to be similar and songs from different classes to be dissimilar, and assessed the relevance of the songs’ rankings returned by each approach.

To assess the relevance of the songs’ rankings, we used the mean average precision (MAP) measure [65]. The MAP is a standard information retrieval measure used in the evaluation of many query-by-example tasks. For each approach and music collection, MAP was computed from the corresponding full distance matrix. The average precision (AP) [65] was computed for each matrix row (for each song query) and the mean was calculated across queries (columns).

For consistency, we applied the same procedure to each of the considered distances, whether or not they required training: the results for RAND,  $L_2$ -PCA,  $L_2$ -RCA-1,  $L_2$ -RCA-2, 1G-MFCC, TEMPO, and CLAS-based distances, were averaged over five iterations of 3-fold cross-validation. On each iteration, all 17 ground truth collections were split into training and testing sets. For each testing set, the CLAS-based distances

were provided with 14 out of 17 training sets. The G3, ART, and ALB collections were not included as training sets due to the insufficient size of their class samples. In contrast, for each testing set,  $L_2$ -RCA-1, and  $L_2$ -RCA-2 were provided with a single complementary training set belonging to the same collection.

### B. Objective Evaluation Results

The average MAP results are presented in Fig. 2 and Table III. Additionally, the approaches with statistically non-significant difference in MAP performance according to the independent two-sample t-tests are presented in Table IV. These t-tests were conducted to separately compare the performances for each music collection. In the cases that are not reported in Table IV, we found statistically significant differences in MAP performance ( $p < 0.05$ ).

We first see that all considered distances outperform the random baseline (RAND) for most of the music collections. When comparing baseline approaches ( $L_2$ -PCA,  $L_2$ -RCA-1,  $L_2$ -RCA-2, 1G-MFCC), we find 1G-MFCC to perform best on average. Still,  $L_2$ -PCA performs similarly (MHA, MSA, MRE, and MEL) or slightly better for some collections (MAC and RPS). With respect to tempo-related collections, TEMPO performs similarly (RPS) or significantly better (RBL) than baseline approaches. Indeed, it is the best performing distance for the RBL collection. Surprisingly, TEMPO yielded accuracies which are comparable to some of the baseline

TABLE IV  
APPROACHES WITH STATISTICALLY NON-SIGNIFICANT DIFFERENCE IN  
MAP PERFORMANCE ACCORDING TO THE INDEPENDENT TWO-SAMPLE  
T-TESTS. THE  $L_2$ -RCA-2 APPROACH WAS EXCLUDED FROM  
THE ANALYSIS DUE TO TECHNICAL DIFFICULTIES

Collection	Compared approaches	P-value
G3	RAND, TEMPO	0.40
MHA	RAND, $L_2$ -RCA-1	1.00
	$L_2$ -PCA, 1G-MFCC	1.00
	CLAS-Pears, CLAS-Pears- $W_M$	0.37
MSA	$L_2$ -PCA, 1G-MFCC	0.37
	CLAS-Pears, CLAS-Pears- $W_M$	0.50
	RAND, TEMPO	1.00
MAG	RAND, TEMPO	1.00
MRE	RAND, TEMPO	0.33
	$L_2$ -PCA, 1G-MFCC	0.09
	CLAS-Pears, CLAS-Pears- $W_M$	0.37
MPA	CLAS-Pears, CLAS-Pears- $W_M$	0.50
MAC	CLAS-Pears, CLAS-Pears- $W_M$	0.08
MEL	$L_2$ -PCA, 1G-MFCC	1.00
	CLAS-Pears, CLAS-Pears- $W_M$	0.37
	RAND, TEMPO	0.33
ALB	CLAS-Pears, CLAS-Pears- $W_M$	0.33
	$L_2$ -RCA-1, TEMPO	1.00
RPS	$L_2$ -RCA-1, TEMPO	1.00

approaches for music collections not strictly related to rhythm or tempo such as G2, MHA, and MEL. In contrast, no statistically significant difference was found in comparison with the random baseline for the G3, MAG, MRE, and ALB collections. Finally, we saw that classifier-based distances achieved the best accuracies for the majority of the collections. Since all CLAS-based distances (CLAS-Cos, CLAS-Pears, CLAS-Spear, CLAS-Cos-W, CLAS-Pears-W, CLAS-Cos-A) showed comparable accuracies, we only report two examples (CLAS-Pears, CLAS-Pears- $W_M$ ). In particular, CLAS-based distances achieved large accuracy improvements with the G2, G4, MPA, MSA, and MAC collections. In contrast, no improvement was achieved with the ART, ALB, and RBL collections. The distance 1G-MFCC performed best for the ART and ALB collections. We hypothesize that the success of 1G-MFCC for the ART and ALB collections might be due to the well-known ‘‘album effect’’ [24]. This effect implies that, due to production process, songs from the same album share much more timbral characteristics than songs from different albums of the same artist, and, moreover, different artists.

### C. Subjective Evaluation Methodology

In the light of the results of the objective evaluation (Section IV-B), we selected four conceptually different approaches ( $L_2$ -PCA, 1G-MFCC, TEMPO, and CLAS-Pears- $W_M$ ) together with the random baseline (RAND) for the listeners’ subjective evaluation. We designed a web-based survey where registered listeners performed a number of iterations blindly voting for the considered distance measures, assessing the quality of how each distance reflects perceived music similarity. In particular, we evaluated the resulting sets of most similar songs produced by the selected approaches, hereafter referred as ‘‘playlists’’. Such a scenario

is a popular way to assess the quality of music similarity measures [3], [6]. It increases discrimination between approaches in comparison with a pairwise song-to-song evaluation. Moreover, it reflects the common applied context of music similarity measurement, which consists of playlist generation.

During each iteration, the listener was presented with 5 different playlists (one for each measure) generated from the same seed song (Fig. 3). Each playlist consisted of the five nearest-to-the-seed songs. The entire process used an in-house collection of 300K music excerpts (30 s) by 60K artists (five songs/artist) covering a wide range of musical dimensions (different genres, styles, arrangements, geographic locations, and epochs). No playlist contained more than one song from the same artist.

Independently for each playlist, we asked the listeners to provide 1) a playlist similarity rating and 2) a playlist inconsistency boolean answer. For playlist similarity ratings, we used a six-point Likert-type scale (0 corresponding to the lowest similarity, 5 to the highest) to evaluate the appropriateness of the playlist with respect to the seed. Likert-type scales [66] are bipolar scales used as tools-of-the-trade in many disciplines to capture subjective information, such as opinions, agreements, or disagreements with respect to a given issue or question. The two opposing positions occupy the extreme ends of the scale (in our case, low-high similarity of the playlist to the seed), and several ratings are allocated for intermediate positions. We explicitly avoided a ‘‘neutral’’ point in order to increase the discrimination between positive and negative opinions. We did not present examples of playlist inconsistency but they might comprise of speech mixed with music, extremely different tempos, completely opposite feelings or emotions, distant musical genres, etc.

We divided the test into two phases: in the first, 12 seeds and corresponding playlists were shared between all listeners; in the second one, the seeds for each listener (up to a maximum of 21) were randomly selected. Listeners were never informed of this distinction. Additionally, we asked each listener about his musical background, which included musicianship and listening expertise information (each measured in three levels). Altogether we collected playlist similarity ratings, playlist inconsistency indicators, and background information from 12 listeners.<sup>8</sup>

### D. Subjective Evaluation Results

In any experimental situation such as our subjective evaluation, analysis of variance (ANOVA) is the usual methodology employed to assess the effects of one variable (like the similarity computation approach) on another one (such as the similarity rating obtained from listeners). ANOVA provides a statistical test of whether or not the means of several groups (in our case, the ratings obtained using a specific similarity computation approach) are equal. In addition to the effect of the different similarity computation methods, in our evaluation we wanted to know the possible effect of the musicianship and listening experience of the participants. Furthermore, we also wanted to know

<sup>8</sup>Due to confidential reasons, the survey was conducted on a limited closed set of participants, and was unavailable to the general public.



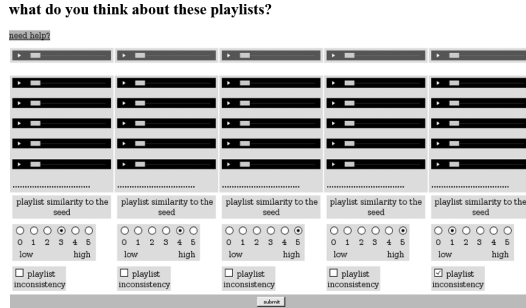


Fig. 3. Screenshot of the subjective evaluation web-based survey.

the effect produced by the two consecutive testing phases used (one presenting the same songs to all the listeners, and the other using different songs for each of them). Therefore, a mixed-design ANOVA with two between-subjects factors (musicianship and listening expertise) and two within-subjects factors (similarity computation approach and testing phase) was required. Results from this analysis revealed that the effect of the similarity computation method on the similarity ratings was statistically significant [Wilks Lambda = 0.005,  $F(4, 2) = 93.943$ ,  $p < 0.05$ ] and that it separated the methods in three different groups: RANDOM and  $L_2$ -PCA (which yielded the lowest similarity ratings) versus TEMPO versus 1G-MFCC and CLAS-Pears- $W_M$  (which yielded the highest similarity ratings). The same pattern was obtained for the effects on the inconsistency ratings (Fig. 4). The effect of the testing phase, also found to be significant, reveals that ratings yielded slightly lower values in the second phase. This could be due to the “tuning” of the similarity ratings experienced by each subject as the experiment proceeded. Fortunately, the impact of phase was uniform and did not depend on or interact with any other factor. Hence, the similarity ratings are only made “finer” or more “selective” as the experiment progresses, but irrespective of the similarity computation approach. On the other hand, the potential effects of musicianship and listening expertise revealed no impact on the similarity ratings. Overall, we conclude that the  $L_2$ -PCA and TEMPO distances, along with a random baseline, revealed poor performance, tending to provide disruptive examples of playlist inconsistency. Contrastingly, CLAS-Pears- $W_M$  and 1G-MFCC revealed acceptable performance with slightly positive user satisfaction. We have omitted for clarity the specific results of the statistical tests which validated our concluding statements.

## V. SEMANTIC EXPLANATION OF MUSIC SIMILARITY

Here we give some thoughts concerning the proposed CLAS distance and its semantic application. An interesting aspect of this proposed approach is the ability to provide a user of the final system with a concrete motivation for the retrieved songs starting from a purely audio content-based analysis. To the best of the authors’ knowledge, this aspect is very rare among other music content-processing systems [67]. However, there is evidence that retrieval or recommendation results perceived

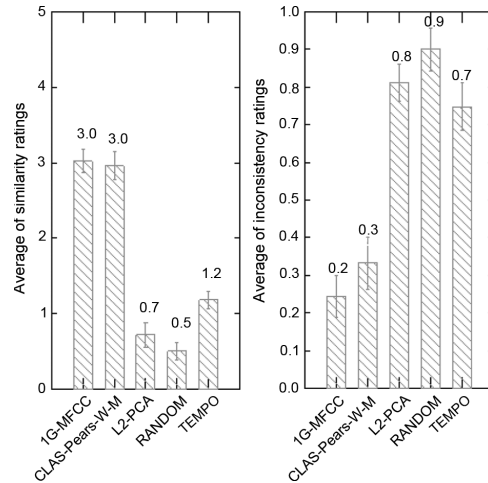


Fig. 4. Average playlist similarity rating and proportion of inconsistent playlists for the subjective evaluation of the simple approaches. Error bars indicate one standard error of the mean.

as transparent (getting an explanation of why a particular retrieval or recommendation was made) are preferred by users, increasing their confidence in a system [68].

Remarkably, the proposed classifier-based distance gives the possibility of providing high-level semantic descriptions for the similarity between a pair of songs along with the distance value itself. In a final system, such annotations can be presented in terms of probability values of the considered dimensions that can be understood by a user. Alternatively, automatic text generation can be employed to present the songs’ qualities in a textual way. For a brief justification of similarity, a subset of dimensions with the highest impact on overall similarity can be selected. A simple use-case example is shown in Fig. 5. For a pair of songs and the CLAS-Pears- $W_M$  distance measure, a subset of 15 dimensions was determined iteratively by greedy distance minimization. In each step, the best candidate for elimination was selected from different dimensions, and its weight was zeroed. Thereafter, the residual dimension probabilities that exceeded corresponding random baselines<sup>9</sup> can be presented to a user. Notice however that as random baselines differ for different dimensions depending on the number of output classes of the corresponding classifier, the significance of dimension probabilities cannot be treated equally. For example, the 0.40 probability of a dimension regressed by an eight-class classifier is considerably more significant than the 0.125 random baseline. Though not presented, the dimensions with probabilities below random baselines also have an impact on the distance measurement. Still, such negative statements (in the sense of a low probability of a regressed dimension) are probably less suitable than positive ones for justification of music similarity to a user.

<sup>9</sup>Under the assumptions of the normal distribution of each classifier’s labels for a music collection.

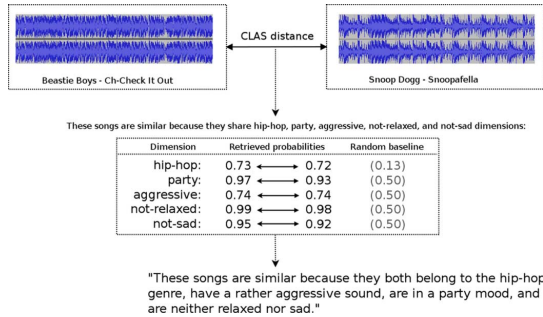


Fig. 5. Real example of a semantic explanation of the similarity between two songs retrieved from our music collection for the classifier-based distance.

## VI. PROPOSED HYBRID APPROACH (HYBRID)

Finally, we hypothesize that an important performance gain can be achieved by combining conceptually different approaches, covering timbral, rhythmic, and semantic aspects of music similarity. We propose a hybrid distance measure, consisting of a subset of the simple measures described above. We define the distance as a weighted linear combination of  $L_2$ -PCA, 1G-MFCC, TEMPO, and CLAS-Pears- $W_M$  distances. We select these four conceptually different approaches relying on the results of the objective evaluation of potential components (Section IV-B). For each selected component, we apply score normalization, following ideas in [69] and [70]. More concretely, each original distance variable  $d_i$  is equalized to a new variable  $\bar{d}_i = E_i(d_i)$ , uniformly distributed in  $[0, 1]$ . The equalizing function  $E_i$  is given by the cumulative distribution function of  $d_i$ , which can be obtained from a distance matrix on a given representative music collection. As such, we use an aggregate collection of 16 000 full songs and music excerpts, composed from the ground truth collections previously used for objective evaluation of simple approaches (Table II). The final hybrid distance is obtained by a weighted linear combination of component distances. The weights are based on the results of the subjective evaluation (Section IV-D) and are set as follows: 0.7 for  $L_2$ -PCA, 3.0 for 1G-MFCC, 1.2 for TEMPO, and 3.0 for CLAS-Pears- $W_M$  distances. Hence, for each component, a weight corresponds to an average playlist similarity rating given by listeners.

## VII. EVALUATION OF HYBRID APPROACH

### A. Objective Evaluation Methodology

Here we followed a different evaluation strategy than with the simple approaches. This strategy comes from the fact that the ground truth music collections available to our evaluation, both in-house and public, can have different biases (due to different collection creators, music availability, audio formats, covered musical dimensions, how the collection was formed, etc.). Therefore, in order to minimize these effects, we carried out a large-scale cross-collection evaluation of the hybrid approach against its component approaches, namely  $L_2$ -PCA, 1G-MFCC, TEMPO, and CLAS-Pears- $W_M$ , together with the random baseline (RAND). Cross-collection comparison implies

TABLE V  
NUMBER OF OCCURRENCES OF TEN MOST FREQUENT GENRES,  
COMMON FOR COLLECTIONS G-C1 AND G-C2

Genre	G-C1	G-C2
Reggae	2991	790
New Age	4294	1034
Blues	6229	2397
Country	8388	1699
Folk	10367	1774
Pop	15796	4523
Electronic	16050	4038
Jazz	22227	5440
Classical	43761	4802
Rock	49369	11486

that the queries and their answers belong to different music collections (out-of-sample results), thus making evaluation results more robust to possible biases.

Solely the genre musical dimension was covered in this experiment. Two large in-house ground truth music collections were employed for that purpose: 1) a collection of 299 000 music excerpts (30 s) (G-C1), and 2) a collection of 73 000 full songs (G-C2). Both collections had a genre label associated with every song. In total, 218 genres and subgenres were covered. The size of these music collections is considerably large, which makes evaluation conditions closer to a real-world scenario. As queries, we randomly selected songs from the ten most common genres from both collections G-C1 and G-C2. The distribution of the selected genres among the collections is presented in Table V. More concretely, for each genre, 790 songs from collection G-C1 were randomly selected as queries. The number of queries per genre corresponds to a minimum number of genre occurrences among the selected genres.

Each query was applied to the collection G-C2, forming a full row in a distance matrix. As with the objective evaluation of simple approaches (Section IV-A), MAP was used as an evaluation measure, but was calculated with a cutoff (similarly to pooling techniques in text retrieval [71]–[73]) equal to the ten closest matches due to the large dimensionality of the resulting distance matrix. The evaluation results were averaged over five iterations. In the same manner, a reverse experiment was carried out, using songs from the G-C2 collection as queries, and applied to the collection G-C1. As the evaluation was completely out-of-sample, the full ground truth collections were used to train the CLAS approach.

### B. Objective Evaluation Results

The results are presented in Table VI. In addition, we analyzed the obtained MAPs with a series of independent two-sample t-tests. All the approaches were found to perform with statistically significant difference ( $p < 0.001$ ).

We see that all considered distances outperform the random baseline (RAND). We found 1G-MFCC and CLAS-Pears- $W_M$  to have comparable performance, being the best among the simple approaches. As well, the TEMPO distance was found to perform similarly or slightly better than  $L_2$ -PCA. Overall, the

TABLE VI  
OBJECTIVE CROSS-COLLECTION EVALUATION RESULTS (MAP  
WITH CUTOFF AT TEN) AVERAGED OVER FIVE ITERATIONS

Distance	G-C1 $\rightarrow$ G-C2	G-C2 $\rightarrow$ G-C1
RANDOM	0.07	0.08
$L_2$ -PCA	0.09	0.11
1G-MFCC	0.23	0.22
TEMPO	0.11	0.12
CLAS-Pears- $W_M$	0.21	0.23
HYBRID	<b>0.25</b>	<b>0.28</b>

results for simple approaches conform with our previous objective evaluation. Meanwhile, our proposed HYBRID distance achieved the best accuracy in the cross-collection evaluation in both directions.

### C. Subjective Evaluation Methodology

We repeated the listening experiment, conducted for simple approaches (Section IV-C) to evaluate the hybrid approach against its component approaches. The same music collection of 300 000 music excerpts (30 s) by 60 000 artists (five songs/artist) was used for that purpose. Each listener was presented with a series of 24 iterations, which, according to the separation of the experiment into two phases, included 12 iterations with seeds and corresponding playlists shared between all listeners, and 12 iterations with randomly selected seeds, different for each listener. In total, we collected playlist similarity ratings, playlist inconsistency indicators, and background information about musicianship and listening expertise from 21 listeners.

### D. Subjective Evaluation Results

An ANOVA with two between-subjects factors (musicianship and listening expertise) and two within-subjects factors (similarity computation approach and testing phase) was used to test their effects on the similarity ratings and on the inconsistency ratings given by the listeners (Fig. 6). The only clearly significant factor explaining the observed variance in the similarity ratings was the similarity computation approach [Wilkslambda = 0.43,  $F(4, 11) = 9.158$ ,  $p < 0.005$ ]. The specific pattern of significant differences between the tested computation approaches makes the HYBRID metric to clearly stand out from the rest, while  $L_2$ -PCA and TEMPO score low (but without statistical differences between them), and CLAS-Pears- $W_M$  and 1G-MFCC (again without statistically significant differences between them) score between the two extremes. As we did not find any significant effect of musicianship and listening expertise on the similarity ratings, it seems clear that the differences in similarity ratings can be attributed only to the differences in the similarity computation approaches.

The same pattern and meaning was also found for the inconsistency ratings: they were dependent on the similarity computation approach, and most of them were generated by the  $L_2$ -PCA and TEMPO methods, whereas the HYBRID method provided

significantly lower inconsistency ratings. No other factor or interaction between factors was found to be statistically significant, but a marginal interaction effect of similarity computation approach and testing phase was found. This effect means that some similarity computation methods (but not all) lowered the ratings as the evaluation progressed. The same pattern was obtained for the inconsistency ratings. In conclusion, we found a similarity computation method (HYBRID) that was clearly preferred over the rest, and no effect other than the computation method was responsible for that preference.

## VIII. MIREX 2009 EVALUATION

### A. Methodology

We submitted the HYBRID and CLAS-Pears- $W_M$  systems to the Music Information Retrieval Evaluation eXchange (MIREX). MIREX is an international community-based framework for the formal evaluation of MIR systems and algorithms [74], [75]. Among other tasks, MIREX allows for the comparison of different algorithms for artist identification, genre classification, or music transcription. In particular, MIREX allows for a subjective human assessment of the accuracy of different approaches to music similarity by community members, this being a central task within the framework. For that purpose, participants can submit their algorithms as binary executables and the MIREX organizers determine and publish the algorithms' accuracies and runtimes. The underlying music collections are never published or disclosed to the participants, neither before or after the contest. Therefore, participants cannot tune their algorithms to the music collections used in the evaluation process.

In the MIREX'2009 edition, the evaluation of each submitted approach was performed on a music collection of 7000 songs (30 s excerpts), which were chosen from IMIRSEL's<sup>10</sup> collections [75] and pertained to ten different genres. For each participant's approach, a  $7000 \times 7000$  distance matrix was calculated. A query set of 100 songs was randomly selected from the music collection, representing each of the ten genres (ten songs per genre). For each query and participant approach, the five nearest-to-the-query songs out of the 7000 were chosen as candidates (after filtering out the query itself and all songs of the same artist). All candidates were evaluated by human graders using the Evalutron 6000 grading system [76]. For each query, a single grader was assigned to evaluate the derived candidates from all approaches. Thereby, the uniformity of scoring within each query was ensured. For each query/candidate pair, a grader provided 1) a categorical broad score in the set  $\{0, 1, 2\}$  (corresponding to "not similar", "somewhat similar", and "very similar" categories), and 2) a fine score in the range from 0 (failure) to 10 (perfection). The listening experiment was conducted with 50 graders, and each one of them evaluated two queries. As this evaluation was completely out-of-sample, our submitted systems were trained on the full ground truth collections required for the CLAS distance.

<sup>10</sup><http://www.music-ir.org/evaluation/>

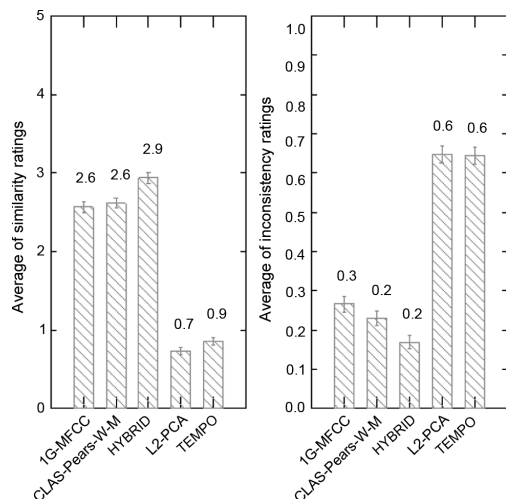


Fig. 6. Average playlist similarity rating and proportion of inconsistent playlists for the subjective evaluation of the hybrid approach. Error bars indicate one standard error of the mean.

## B. Results

The overall evaluation results are reproduced in Table VII.<sup>11</sup> Our measures are noted as BSWH1 for CLAS-Pears- $W_M$ , and BSWH2 for HYBRID. The results of the Friedman test against the summary data of fine scores are presented in Fig. 7. First, and most importantly, we found the HYBRID measure to be one of the best performing distances in the MIREX 2009 audio music similarity task. HYBRID was very close to PS1, but worse than the leading PS2 distance [15]. However, no statistically significant difference between PS2, PS1, and our HYBRID measure was found in the Friedman test. Second, the CLAS-Pears- $W_M$  measure revealed satisfactory average performance comparing to other distances with no statistically significant difference to the majority of the participant approaches. Nevertheless, CLAS-Pears- $W_M$  outperformed a large group of poor performing distances with a statistically significant difference. Finally, we state that despite the fact that we do not observe examples of stable excellent performance among all participant distances, up to above-average user satisfaction was achieved by the majority of the approaches, including our HYBRID and CLAS-Pears- $W_M$  distances.

## IX. CONCLUSIONS

In the current work, we presented, studied, and comprehensively evaluated, both objectively and subjectively, new and existing content-based distance measures for music similarity. We studied a number of simple approaches, each of which apply a uniform distance measure for overall similarity. We considered five baseline distances, including a random one. We explored the potential of two new conceptually different distances not strictly

<sup>11</sup>Detailed results can be found on the official results webpage for MIREX'2009: [http://www.music-ir.org/mirex/2009/index.php/Audio\\_Music\\_Similarity\\_and\\_Retrieval\\_Results](http://www.music-ir.org/mirex/2009/index.php/Audio_Music_Similarity_and_Retrieval_Results)

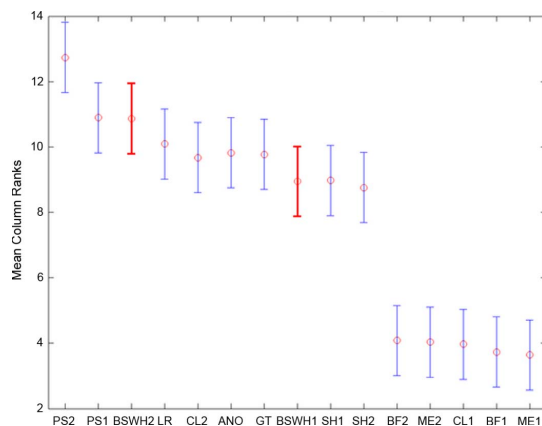


Fig. 7. MIREX 2009 Friedman's test (fine scores). Figure obtained from the official results webpage for MIREX'2009.

operating on the often exclusively used musical timbre aspects. More concretely, we presented a simple tempo-based distance which can be especially useful for expressing music similarity in collections where rhythm aspects are predominant. Using only two low-level temporal descriptors, BPM and OR, this distance is computationally inexpensive, yet effective for such collections. As well, our subjective evaluation experiments revealed a slight preference by listeners of tempo-based distance over a generic Euclidean distance.

In addition, we investigated the possibility of benefiting from the results of classification problems and transferring this gained knowledge to the context of music similarity. To this end, we presented a classifier-based distance which makes use of high-level semantic descriptors inferred from low-level ones. This distance covers diverse groups of musical dimensions such as genre and musical culture, moods and instruments, and rhythm and tempo. The classifier-based distance outperformed all the considered simple approaches in most of the ground truth music collections used for objective evaluation. In contrast, this performance improvement was not seen in the subjective evaluation when compared with the best performing baseline distance considered. However, they were found to perform at the same level and, therefore, no statistically significant differences were found between them. In general, the classifier-based distance represents a semantically rich approach to music similarity. Thus, in spite of being based solely on audio content information, this approach can overcome the so-called "semantic gap" in content-based music similarity and provide a semantic explanation to justify the retrieval results to a user.

We explored the possibility of creating a hybrid approach, based on the studied simple approaches as potential components. We presented a new distance measure, which combines a low-level Euclidean distance based on PCA, a timbral distance based on single Gaussian MFCC modeling, our tempo-based distance, and a high-level semantic classifier-based distance. This distance outperformed all previously considered approaches in an objective large-scale cross-collection evaluation, and revealed the best performance for listeners in a subjective

TABLE VII  
MIREX 2009 OVERALL SUMMARY RESULTS SORTED BY AVERAGE FINE SCORE. THE PROPOSED APPROACHES  
CLAS AND HYBRID ARE HIGHLIGHTED IN GRAY (BSWH1 AND BSWH2, RESPECTIVELY)

Acronym	Authors (measure)	Average fine score	Average broad score
PS2	Tim Pohle, Dominik Schnitzer (2009)	6.458	1.448
PS1	Tim Pohle, Dominik Schnitzer (2007)	5.751	1.262
BSWH2	Dmitry Bogdanov, Joan Serrà, Nicolas Wack, and Perfecto Herrera (HYBRID)	5.734	1.232
LR	Thomas Lidy, Andreas Rauber	5.470	1.148
CL2	Chuan Cao, Ming Li	5.392	1.164
ANO	Anonymous	5.391	1.126
GT	George Tzanetakis	5.343	1.126
BSWH1	Dmitry Bogdanov, Joan Serrà, Nicolas Wack, and Perfecto Herrera (CLAS-Pears- $W_M$ )	5.137	1.094
SH1	Stephan Hübler	5.042	1.012
SH2	Stephan Hübler	4.932	1.040
BF2	Benjamin Fields (mfcc10)	2.587	0.410
ME2	François Maillat, Douglas Eck (sda)	2.585	0.418
CL1	Chuan Cao, Ming Li	2.525	0.476
BF1	Benjamin Fields (chr12)	2.401	0.416
ME1	François Maillat, Douglas Eck (mlp)	2.331	0.356

evaluation. Moreover, we participated in a subjective evaluation against a number of state-of-the-art distance measures, within the bounds of the MIREX'2009 audio music similarity and retrieval task. The results revealed high performance of our hybrid measure, with no statistically significant difference from the best performing method submitted. In general, the hybrid distance represents a combinative approach, benefiting from timbral, rhythmic, and high-level semantic aspects of music similarity.

Further research will be devoted to improving the classifier-based distance with the addition of classifiers dealing with musical dimensions such as tonality or instrument information. Given that several separate dimensions can be straightforwardly combined with this distance, additional improvements are feasible and potentially beneficial. In particular, contextual dimensions, in the form of user ratings or social tags, can be added to make possible a fusion with collaborative filtering approaches. As well, to improve the classifier-based distance itself, we will consider a better combination of classifiers' output probabilities. Additionally, an enhancement of the tempo-based distance component of the proposed hybrid approach is possible by using a richer representation for rhythm, such as the fluctuation patterns.

#### ACKNOWLEDGMENT

The authors would like to thank J. Funollet for technical support, O. Meyers for technical support and proofreading, and all participants of the subjective evaluation.

#### REFERENCES

- [1] G. Lu, "Techniques and data structures for efficient multimedia retrieval based on similarity," *IEEE Trans. Multimedia*, vol. 4, no. 3, pp. 372–384, 2002.
- [2] M. A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney, "Content-based music information retrieval: Current directions and future challenges," *Proc. IEEE*, vol. 96, no. 4, pp. 668–696, 2008.
- [3] M. Slaney and W. White, "Similarity based on rating data," in *Proc. Int. Symp. Music Information Retrieval (ISMIR'07)*, 2007.
- [4] M. Levy and M. Sandler, "Music information retrieval using social tags and audio," *IEEE Trans. Multimedia*, vol. 11, no. 3, pp. 383–395, 2009.

- [5] O. Celma, "Music recommendation and discovery in the long tail," Ph.D. dissertation, Univ. Pompeu Fabra, Barcelona, Spain, 2008.
- [6] L. Barrington, R. Oda, and G. Lanckriet, "Smarter than genius? human evaluation of music recommender systems," in *Proc. International Society for Music Information Retrieval Conf. (ISMIR'09)*, 2009, pp. 357–362.
- [7] D. Bogdanov, J. Serrà, N. Wack, and P. Herrera, "From low-level to high-level: Comparative study of music similarity measures," in *Proc. IEEE Int. Symp. Multimedia (ISM'09). Int. Workshop Advances in Music Information Research (AdMIR'09)*, 2009, pp. 453–458.
- [8] L. Barrington, D. Turnbull, D. Torres, and G. Lanckriet, "Semantic similarity for music retrieval," in *Proc. Music Information Retrieval Evaluation Exchange (MIREX'07)*, 2007. [Online]. Available: [http://www.music-ir.org/mirex/abstracts/2007/AS\\_barrington.pdf](http://www.music-ir.org/mirex/abstracts/2007/AS_barrington.pdf).
- [9] A. Berenzweig, D. P. W. Ellis, and S. Lawrence, "Anchor space for classification and similarity measurement of music," in *Proc. Int. Conf. Multimedia and Expo (ICME'03)*, 2003, vol. 1, pp. 29–32.
- [10] K. West and P. Lamere, "A model-based approach to constructing music similarity functions," *EURASIP J. Adv. Signal Process.*, vol. 2007, p. 149, 2007.
- [11] G. C. Cupchik, M. Rickert, and J. Mendelson, "Similarity and preference judgments of musical stimuli," *Scandinavian J. Psychol.*, vol. 23, no. 4, pp. 273–282, 1982.
- [12] E. Pampalk, A. Flexer, and G. Widmer, "Improvements of audio-based music similarity and genre classification," in *Proc. Int. Conf. Music Information Retrieval (ISMIR'05)*, 2005, pp. 628–633.
- [13] E. Pampalk, "Computational models of music similarity and their application in music information retrieval," Ph.D. dissertation, Vienna Univ. Technol., Vienna, Austria, 2006.
- [14] T. Pohle and D. Schnitzer, "Striving for an improved audio similarity measure," in *Proc. Music Information Retrieval Evaluation Exchange (MIREX'07)*, 2007. [Online]. Available: [http://www.music-ir.org/mirex/2007/abs/AS\\_pohle.pdf](http://www.music-ir.org/mirex/2007/abs/AS_pohle.pdf).
- [15] T. Pohle and D. Schnitzer, "Submission to MIREX 2009 audio similarity task," in *Proc. Music Information Retrieval Evaluation Exchange (MIREX'09)*, 2009. [Online]. Available: <http://music-ir.org/mirex/2009/results/abs/PS.pdf>.
- [16] T. Pohle, D. Schnitzer, M. Schedl, P. Knees, and G. Widmer, "On rhythm and general music similarity," in *Proc. International Society for Music Information Retrieval Conf. (ISMIR'09)*, 2009, pp. 525–530.
- [17] Y. Song and C. Zhang, "Content-based information fusion for semi-supervised music genre classification," *IEEE Trans. Multimedia*, vol. 10, no. 1, pp. 145–152, 2008.
- [18] B. McFee and G. Lanckriet, "Heterogeneous embedding for subjective artist similarity," in *Proc. Int. Conf. Music Information Retrieval (ISMIR'09)*, 2009.
- [19] P. Cano, M. Koppenberger, and N. Wack, "Content-based music audio recommendation," in *Proc. ACM Int. Conf. Multimedia (ACMMM'05)*, 2005, pp. 211–212.

- [20] M. Slaney, K. Weinberger, and W. White, "Learning a metric for music similarity," in *Proc. Int. Symp. Music Information Retrieval (ISMIR'08)*, 2008, pp. 313–318.
- [21] N. Shental, T. Hertz, D. Weinshall, and M. Pavel, "Adjustment learning and relevant component analysis," *Lecture Notes In Computer Science*, pp. 776–792, 2002.
- [22] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learn. Res.*, vol. 10, pp. 207–244, 2009.
- [23] B. Logan and A. Salomon, "A music similarity function based on signal analysis," in *Proc. IEEE Int. Conf. Multimedia and Expo (ICME'01)*, 2001, p. 190.
- [24] M. I. Mandel and D. P. Ellis, "Song-level features and support vector machines for music classification," in *Proc. Int. Conf. Music Information Retrieval (ISMIR'05)*, 2005, pp. 594–599.
- [25] J. J. Aucouturier and F. Pachet, "Music similarity measures: What's the use," in *Proc. ISMIR*, 2002, pp. 157–163.
- [26] J. J. Aucouturier, F. Pachet, and M. Sandler, "The way it sounds: Timbre models for analysis and retrieval of music signals," *IEEE Trans. Multimedia*, vol. 7, no. 6, pp. 1028–1035, 2005.
- [27] A. Flexer, D. Schnitzer, M. Gasser, and G. Widmer, "Playlist generation using start and end songs," in *Proc. Int. Symp. Music Information Retrieval (ISMIR'08)*, 2008, pp. 173–178.
- [28] J. H. Jensen, M. G. Christensen, D. P. W. Ellis, and S. H. Jensen, "Quantitative analysis of a common audio similarity measure," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, pp. 693–703, 2009.
- [29] T. Li and M. Ogihara, "Toward intelligent music information retrieval," *IEEE Trans. Multimedia*, vol. 8, no. 3, pp. 564–574, 2006.
- [30] D. P. Ellis and G. E. Poliner, "Identifying 'cover songs' with chroma features and dynamic programming beat tracking," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP'07)*, 2007, pp. 1429–1432.
- [31] M. Marolt, "A Mid-level representation for melody-based retrieval in audio collections," *IEEE Trans. Multimedia*, vol. 10, no. 8, pp. 1617–1625, 2008.
- [32] J. Serra, X. Serra, and R. G. Andrzejak, "Cross recurrence quantification for cover song identification," *New J. Phys.*, vol. 11, no. 9, p. 093017, 2009.
- [33] O. Celma, P. Herrera, and X. Serra, "Bridging the music semantic gap," in *Proc. ESWC 2006 Workshop Mastering the Gap: From Information Extraction to Semantic Representation*, 2006. [Online]. Available: <http://mtg.upf.edu/node/874>.
- [34] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293–302, 2002.
- [35] C. Laurier, O. Meyers, J. Serrà, M. Blech, P. Herrera, and X. Serra, "Indexing music by mood: Design and integration of an automatic content-based annotator," *Multimedia Tools and Applications*, vol. 48, no. 1, pp. 161–184, 2010.
- [36] C. Laurier, O. Meyers, J. Serrà, M. Blech, and P. Herrera, "Music mood annotator design and integration," in *Proc. Int. Workshop Content-Based Multimedia Indexing (CBMI'2009)*, 2009.
- [37] G. Peeters, A Large Set of Audio Features for Sound Description (Similarity and Classification) in the CUIDADO Project CUIDADO Project Report, 2004. [Online]. Available: <http://recherche.ircam.fr/equipes/analyse-synthese/peeters/ARTICLES/>.
- [38] B. Logan, "Mel frequency cepstral coefficients for music modeling," in *Proc. Int. Symp. Music Information Retrieval (ISMIR'00)*, 2000.
- [39] P. M. Brossier, "Automatic annotation of musical audio for interactive applications," Ph.D. dissertation, Queen Mary, Univ. London, London, U.K., 2007.
- [40] E. Gómez, P. Herrera, P. Cano, J. Janer, J. Serrà, J. Bonada, S. El-Hajj, T. Aussenac, and G. Holmberg, "Music Similarity Systems and Methods Using descriptors," WO 2009/001202, Dec. 31, 2008.
- [41] F. Gouyon, "A computational approach to rhythm description: Audio features for the computation of rhythm periodicity functions and their use in tempo induction and music content processing," Ph.D. dissertation, Univ. Pompeu Fabra, Barcelona, Spain, 2005.
- [42] E. Gómez, "Tonal description of music audio signals," Ph.D. dissertation, Univ. Pompeu Fabra, Barcelona, Spain, 2006.
- [43] W. A. Sethares, *Tuning, Timbre, Spectrum, Scale*. New York: Springer Verlag, 2005.
- [44] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco, CA: Morgan Kaufmann, 2005.
- [45] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is 'nearest neighbor' meaningful?," in *Proc. Database Theory—ICDT'99*, 1999, pp. 217–235.
- [46] F. Korn, B. Pagel, and C. Faloutsos, "On the 'dimensionality curse' and the 'self-similarity blessing,'" *IEEE Trans. Knowl. Data Eng.*, vol. 13, no. 1, pp. 96–111, 2001.
- [47] C. C. Aggarwal, "On k-anonymity and the curse of dimensionality," in *Proc. Int. Conf. Very Large Data Bases (VLDB'05)*, VLDB Endowment, Trondheim, Norway, 2005, pp. 901–909.
- [48] N. Wack, P. Cano, B. de Jong, and R. Marxer, *A Comparative Study of Dimensionality Reduction Methods: The Case of Music Similarity*, Music Technology Group, 2006, Tech. Rep. [Online]. Available: [http://mtg.upf.edu/files/publications/NWack\\_2006.pdf](http://mtg.upf.edu/files/publications/NWack_2006.pdf).
- [49] T. Pohle, P. Knees, M. Schedl, and G. Widmer, "Automatically adapting the structure of audio similarity spaces," in *Proc. Workshop Learning the Semantics of Audio Signals (LSAS'06)*, 2006, pp. 66–75.
- [50] E. Pampalk, S. Dixon, and G. Widmer, "On the evaluation of perceptual similarity measures for music," in *Proc. 6th Int. Conf. Digital Audio Effects (DAFx'03)*, London, U.K., 2003, pp. 7–12.
- [51] S. Sigurdsson, K. B. Petersen, and T. Lehn-Schiöler, "Mel frequency cepstral coefficients: An evaluation of robustness of mp3 encoded music," in *Proc. Int. Conf. Music Information Retrieval (ISMIR'07)*, 2006, pp. 286–289.
- [52] M. F. McKinney and D. Moelants, "Ambiguity in tempo perception: What draws listeners to different metrical levels?," *Music Percept.*, vol. 24, no. 2, pp. 155–166, 2006.
- [53] F. Gouyon, A. Klapuri, S. Dixon, M. Alonso, G. Tzanetakis, C. Uhle, and P. Cano, "An experimental comparison of audio tempo induction algorithms," *IEEE Trans. Speech Audio Process.*, vol. 14, no. 5, pp. 1832–1844, 2006.
- [54] L. M. Smith, "Beat critic: Beat tracking octave error identification by metrical profile analysis," in *Proc. International Society for Music Information Retrieval Conf. (ISMIR'10)*, 2010.
- [55] E. Gómez and P. Herrera, "Comparative analysis of music recordings from western and non-western traditions by automatic tonal feature extraction," *Empiric. Musicol. Rev.*, vol. 3, no. 3, pp. 140–156, 2008.
- [56] C. Xu, N. C. Maddage, X. Shao, F. Cao, and Q. Tian, "Musical genre classification using support vector machines," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'03)*, 2003, pp. 429–432.
- [57] B. Sarwar, G. Karypis, J. Konstan, and J. Reidl, "Item-based collaborative filtering recommendation algorithms," in *Proc. Int. Conf. World Wide Web (WWW'01)*, 2001, pp. 285–295.
- [58] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl, "Evaluating collaborative filtering recommender systems," *ACM Trans. Inf. Syst.*, vol. 22, no. 1, pp. 5–53, 2004.
- [59] A. Cripps, C. Pettey, and N. Nguyen, "Improving the performance of FLN by using similarity measures and evolutionary algorithms," in *Proc. IEEE Int. Conf. Fuzzy Systems*, 2006, pp. 323–330.
- [60] M. B. Abdullah, "On a robust correlation coefficient," *J. R. Statist. Soc. Series D (The Statistician)*, vol. 39, no. 4, pp. 455–460, 1990.
- [61] A. Novello, M. F. McKinney, and A. Kohlrausch, "Perceptual evaluation of music similarity," in *Proc. Int. Conf. Music Information Retrieval (ISMIR'06)*, 2006.
- [62] H. Homburg, I. Mierswa, B. Möller, K. Morik, and M. Wurst, "A benchmark dataset for audio classification and clustering," in *Proc. Int. Conf. Music Information Retrieval (ISMIR'05)*, 2005, pp. 528–531.
- [63] P. J. Rentfrow and S. D. Gosling, "The do re mi's of everyday life: The structure and personality correlates of music preferences," *J. Personal. Social Psychol.*, vol. 84, pp. 1236–1256, 2003.
- [64] P. Cano, E. Gómez, F. Gouyon, P. Herrera, M. Koppenberger, B. Ong, X. Serra, S. Streich, and N. Wack, ISMIR 2004 Audio Description Contest, 2006. [Online]. Available: <http://mtg.upf.edu/node/461>.
- [65] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge Univ. Press, 2008.
- [66] W. E. Saris and I. N. Gallhofer, *Design, Evaluation and Analysis of Questionnaires for Survey Research*. New York: Wiley-Interscience, 2007.
- [67] F. Mailet, D. Eck, G. Desjardins, and P. Lamere, "Steerable playlist generation by learning song similarity from radio station playlists," in *Proc. Int. Conf. Music Information Retrieval (ISMIR'09)*, 2009.
- [68] R. Sinha and K. Swearingen, "The role of transparency in recommender systems," in *Proc. CHI'02 Extended Abstracts on Human Factors in Computing Systems*, 2002, p. 831.
- [69] M. Fernández, D. Vallet, and P. Castells, "Probabilistic score normalization for rank aggregation," in *Proc. Advances in Information Retrieval*, 2006, pp. 553–556.
- [70] M. Arevalillo-Herráez, J. Domingo, and F. J. Ferri, "Combining similarity measures in content-based image retrieval," *Pattern Recognit. Lett.*, vol. 29, no. 16, pp. 2174–2181, Dec. 2008.

- [71] A. Turpin and F. Scholer, "User performance versus precision measures for simple search tasks," in *Proc. Int. ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR'06)*, 2006, pp. 11–18.
- [72] W. B. Croft, D. Metzler, and T. Strohman, *Search Engines: Information Retrieval in Practice*. Reading, MA: Addison-Wesley, 2010.
- [73] F. Radlinski and N. Craswell, "Comparing the sensitivity of information retrieval metrics," in *Proc. 33rd Int. ACM SIGIR Conf. Research and Development in Information Retrieval*, 2010, pp. 667–674.
- [74] J. S. Downie, "The music information retrieval evaluation exchange (2005–2007): a window into music information retrieval research," *Acoust. Sci. Technol.*, vol. 29, no. 4, pp. 247–255, 2008.
- [75] J. Downie, A. Ehmann, M. Bay, and M. Jones, "The music information retrieval evaluation eXchange: some observations and insights," in *Proc. Advances in Music Information Retrieval*, 2010, pp. 93–115.
- [76] A. A. Grudz, J. S. Downie, M. C. Jones, and J. H. Lee, "Evalutron 6000: Collecting music relevance judgments," in *Proc. ACM/IEEE-CS Joint Conf. Digital Libraries (JCDL'07)*, 2007, p. 507.



**Dmitry Bogdanov** received a degree in applied mathematics and informatics at the Lomonosov Moscow State University, Moscow, Russia, in 2006. He is presently pursuing the Ph.D. degree at the Universitat Pompeu Fabra (UPF), Barcelona, Spain.

While affiliated with MSU, he participated in several research projects devoted to network security and intrusion detection. In 2007, he became a member of the Music Technology Group at UPF as a researcher. His current research interests include music information retrieval, music similarity and recommendation,

music content description and classification, user modeling, and preference elicitation.



**Joan Serrà** received the telecommunications and electronics degrees at Enginyeria La Salle, Universitat Ramon Llull, Barcelona, Spain, in 2002 and 2004, respectively. After working from 2005 to 2006 at the research and development department of Music Intelligence Solutions Inc., he joined the Music Technology Group of Universitat Pompeu Fabra, Barcelona, where he received the M.Sc. degree and the Ph.D. degree in information, communication, and audiovisual media technologies in 2007 and 2011, respectively.

He is currently a post-doc researcher with the Music Technology Group of the UPF. He is also a part-time associate professor with the Dept. of Information and Communication Technologies of the same university. In 2010, he was a guest scientist with the Research Group on Nonlinear Dynamics and Time Series Analysis of the Max Planck Institute for the Physics of Complex Systems in Dresden, Germany. His main research interests include music retrieval and understanding, signal processing, time series analysis, complex networks, complex systems, information retrieval and music perception, psychology, and cognition.



**Nicolas Wack** received the Telecommunication Engineer degree from Telecom ParisTech, Paris, France, in 2003.

Since then, he has been involved with the Music Technology Group, Universitat Pompeu Fabra (UPF), Barcelona, Spain, in various European projects dealing with sound and music classification and similarity. He spearheaded the development of the *Essentia* and *Gaia* technologies, which, respectively, extract audio characteristics from sounds/songs and perform similarity queries on

them. His main interests are in software design, efficient audio analysis, and working with very large databases.



**Perfecto Herrera** received a degree in psychology from the University of Barcelona, Barcelona, Spain, in 1987. Now he is pursuing the Ph.D. in music content description in the Universitat Pompeu Fabra (UPF), Barcelona, Spain.

He was with the University of Barcelona as a Software Developer and as Assistant Professor. His further studies focused on sound engineering, audio postproduction, and computer music. He has been working with the Music Technology Group, UPF, since its inception in 1996, first as the person responsible for the sound laboratory/studio, then as a Researcher. He worked in the MPEG-7 standardization initiative from 1999 to 2001. Then, he collaborated in the EU-IST-funded CUIDADO project, contributing to the research and development of tools for indexing and retrieving music and sound collections. This work continued and was expanded as Scientific Coordinator for the Semantic Interaction with Music Audio Contents (SIMAC) project, again funded by the EU-IST. He is currently the Head of the Department of Sonology, Higher Music School of Catalonia (ESMUC), where he teaches music technology and psychoacoustics. His main research interests are music content analysis, description and classification, and music perception and cognition.

able for the sound laboratory/studio, then as a Researcher. He worked in the MPEG-7 standardization initiative from 1999 to 2001. Then, he collaborated in the EU-IST-funded CUIDADO project, contributing to the research and development of tools for indexing and retrieving music and sound collections. This work continued and was expanded as Scientific Coordinator for the Semantic Interaction with Music Audio Contents (SIMAC) project, again funded by the EU-IST. He is currently the Head of the Department of Sonology, Higher Music School of Catalonia (ESMUC), where he teaches music technology and psychoacoustics. His main research interests are music content analysis, description and classification, and music perception and cognition.



**Xavier Serra** received the Ph.D. degree in computer music from Stanford University, Stanford, CA, in 1989, with a dissertation on the spectral processing of musical sounds that is considered a key reference in the field.

He is an Associate Professor of the Department of Information and Communication Technologies and Director of the Music Technology Group at the Universitat Pompeu Fabra, Barcelona, Spain. His research interests cover the understanding, modeling, and generation of musical signals by computational

means, with a balance between basic and applied research and approaches from both scientific/technological and humanistic/artistic disciplines. He is very active in promoting initiatives in the field of Sound and Music Computing at the local and international levels, being editor and reviewer of a number of journals, conferences, and research programs of the European Commission, and also giving lectures on current and future challenges of the field. He is the principal investigator of more than 15 major research projects funded by public and private institutions, the author of 31 patents and of more than 75 research publications.

Cano, P., Koppenberger, M., Le Groux, S., Ricard, J., Wack, N., **Herrera, P.** (2005). "Nearest-neighbor sound annotation with a Wordnet taxonomy". *Journal of Intelligent Information Systems*, 24 (2-3), pp. 99-111.

DOI: <https://doi.org/10.1007/s10844-005-0318-4>

ISSN 0925-9902

Online ISSN 1573-7675



# Nearest-neighbor Automatic Sound Annotation with a WordNet Taxonomy

Pedro Cano, Markus Koppenberger, Sylvain Le Groux,  
Julien Ricard, Nicolas Wack and Perfecto Herrera  
Music Technology Group, Institut de l'Audiovisual  
Universitat Pompeu Fabra, 08003 Barcelona, Spain  
Phone: +34 935 422 101  
Fax: +34 935 422 202  
E-mail: pcano@iua.upf.es  
Web: www.iua.upf.es/mtg

October 15, 2004

## Abstract

Sound engineers need to access vast collections of sound effects for their film and video productions. Sound effects providers rely on text-retrieval techniques to offer their collections. Currently, annotation of audio content is done manually, which is an arduous task. Automatic annotation methods, normally fine-tuned to reduced domains such as musical instruments or reduced sound effects taxonomies, are not mature enough for labeling with great detail any possible sound. A general sound recognition tool would require: first, a taxonomy that represents the world and, second, thousands of classifiers, each specialized in distinguishing little details. We report experimental results on a general sound annotator. To tackle the taxonomy definition problem we use WordNet, a semantic network that organizes real world knowledge. In order to overcome the need of a huge number of classifiers to distinguish many different sound classes, we use a nearest-neighbor classifier with a database of isolated sounds unambiguously linked to WordNet concepts. A 30% concept prediction is achieved on a database of over 50.000 sounds and over 1600 concepts.

## 1 INTRODUCTION

Sound effects providers rely on classical text retrieval techniques to manage manually labeled audio collections. The manual annotation is a labor-intensive and error-prone task. There are attempts towards metadata generation by automatic classification. State of the art of audio classification methods, except for reduced-domain tasks, is not mature enough for real world applications. Audio classification methods cannot currently provide the level of detail needed in a sound effect management system, e.g: “fast female footsteps on wood”, “violin pizzicato with natural open strings” or “mid tom with loose skin bend at end”. In audio classification, researchers normally assume the existence of

a well defined hierarchical classification scheme of a few categories (less than a hundred categories). On-line sound effects and music sample providers have several thousand categories. This makes the idea of generating a model for each category quite unfeasible, we would need several thousand classifiers.

In this context, we present an all-purpose sound recognition system based on nearest-neighbor classification rule. A sound sample will be labeled with the descriptions from the similar sounding examples of a annotated database. The terms borrowed from the closest match are unambiguous due to the use of WordNet<sup>1</sup> (Miller November 1995) as the taxonomy back-end. With unambiguous tagging, we refer to assigning concepts and not just terms to sounds. For instance, the sound of a “bar” is ambiguous, it could be “bar” as “rigid piece of metal or wood” or as “establishment where alcoholic drinks are served” where each concept has a unique identifier.

The rest of the paper is organized as follows: In Section 2 we briefly enumerate some approaches to the problem of automatically identification and we reflect on the difficulties inherent in automatically describing any isolated sounds with a high level of detail. In Section 3, we present a real world size taxonomy for sound effect description. From Section 4 to 7 we describe the system setup as well as preliminary results. We conclude with possible continuations of the approach.

## 2 RELATED WORK

Existing classification methods are normally finely tuned to small domains, such as musical instrument classification (Herrera, et al. 2003)(Kostek & Czyzewski 2001) or simplified sound effects taxonomies (Casey 2002)(Wold, et al. 1996), (Zhang & Kuo 1999). Peltonen *et al.* presented a system in (Peltonen, et al. 2002) devised to classify environments or ambiances, e.g: “street, pub, office, church”. Different audio classification systems differ mainly on the acoustic features derived from the sound and the type of classifier. Independently of the feature extraction and selection method and the type of classifier used, content-based classification systems need a set of classes and a large number (e.g: 30 or more) of audio samples for each class to train the system.

Classification methods cannot currently offer the detail needed in commercial sound effects management. It would require to develop thousands of classifiers, each specialized in distinguishing little details and a taxonomy that represents the real world. Dubnov and Ben-Shalom (Dubnov & Ben-Shalom 2003) point that one of the main problems faced by natural sounds and sound effects classifiers is the lack of clear taxonomy. In musical instrument classification, the taxonomies more or less follow perceptual-related hierarchical structures(Lakatos 2000). Accordingly, in such problems one can devise hierarchical classification approaches such as (Martin 1999)(Peeters & Rodet 2003) in which the system distinguishes in a first level between sustained and non-sustained sounds, and in a second level among strings, woodwinds and so on. In every-day sound classification, there is not such a parallelism between semantic and perceptual categories. On the contrary one can find hissing sounds in categories of “cat”, “tea boilers”, “snakes”. Sound engineers exploit this ambiguity and create the illusion of “crackling fire” by “recording twisting cellophane”.

---

<sup>1</sup><http://www.cogsci.princeton.edu/~wn/>

We have to add to this problem the fact that designing a taxonomy or classification scheme to include the concepts of the real world is a daunting task. The MPEG7 standard provides description mechanisms and taxonomy management tools for multimedia documents. Casey (Casey 2002) shows an example on how to build such a classification scheme using MPEG7. However, it is very complicated to devise and maintain classification schemes that account for the level of detail needed in a production-size sound effect management system. We have found that it is much faster to start developing ontologies on top of a semantic network such as WordNet rather than starting from scratch (see Section 3).

Slaney describes in (Slaney 2002) a method of connecting words to sounds. He avoids the needs of taxonomy design when bridging the gap between perceptual and semantic spaces searching for hierarchies in an unsupervised mode. Barnard *et al.* describe a similar approach for matching words and images (Barnard, et al. 2003).

### 3 TAXONOMY MANAGEMENT

WordNet is a lexical network designed following psycholinguistic theories of human lexical memory. Standard dictionary organize words alphabetically. WordNet organizes concepts in synonym sets, called *synsets*, with links between the concepts. It knows for instance that the word piano, as a noun, has two senses, the musical attribute that refers to “low loudness” and the “musical instrument”. It also encodes the information that a “grand piano” is a type of “piano”, and that it has parts such as a keyboard, a loud pedal and so on. Such a knowledge system is useful for retrieval. It can for instance display the results of a query “car” in types of cars, parts of car, actions of a car (approaching, departing, turning off).

Even though WordNet already organizes over 100,000 terms, it sometimes lacks specific knowledge, such as “close-up” — referring to the recording technique — or that a “747” is an airplane. We have developed a WordNet editor and augmented it both with concepts from taxonomies to describe acoustically sounds and mining legacy metadata from sound effects libraries. The extended lexical network includes the semantic aspects, perceptual and sound effects specific terms in an unambiguous way. For further details on the implementation and evaluation of WordNet as backbone for audio taxonomy management, we refer to (Cano, et al. 2004b).

### 4 EXPERIMENTAL SETUP

Our database consists of 54,799 sounds from over 30 different libraries of sound effects, music and music samples. These sounds have been unambiguously tagged with concepts of an enhanced WordNet. Thus a violin sound with the following caption: “violin pizzicato D#” has the following synsets:

- violin, fiddle – (bowed stringed instrument that is the highest member of the violin family; this instrument has four strings and a hollow body and an unfretted fingerboard and is played with a bow)

- pizzicato – ((of instruments in the violin family) to be plucked with the finger)
- re, ray – (the syllable naming the second (supertonic) note of any major scale in solmization)
- sharp – ((music) raised in pitch by one chromatic semitone; "C sharp")

In Figure 1, we show a histogram with the number of synsets the sounds have been labeled with after disambiguation. It should be clear that the higher the number of synsets, the better a sound is described. In average, a sound is labeled with 3.88 synsets. In Figure 2 we plot the rank-frequency analysis of the synsets. For this analysis we counted the occurrence of different synsets and then sorted them according to descending frequency. The plot is repeated for various parts of speech, specifically: noun, verb, adjective and adverb. The distribution of 3028 synsets with respect its syntactic function is as follows: 2381 nouns, 380 verbs, 251 adjectives and 16 adverbs. The number of synsets for which there are ten or more examples sounds is 1645.

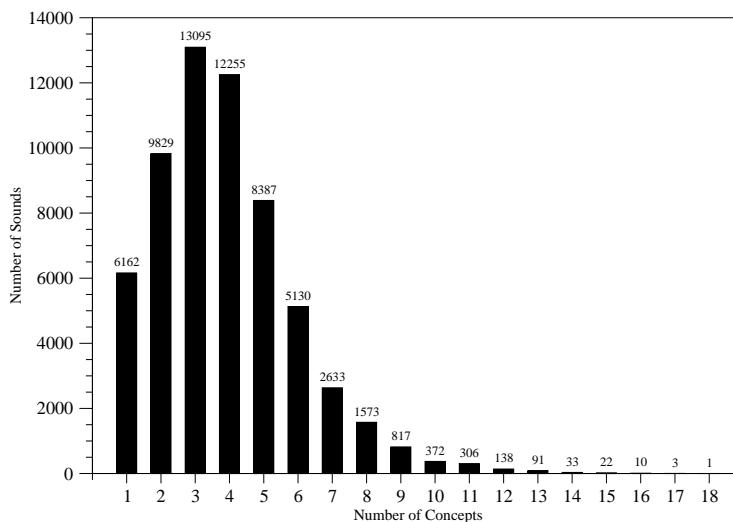


Figure 1: Histogram of number of synsets (concepts) per sound

The classifier uses a set of 89 features and a nearest-neighbor classifier using a database of sounds with WordNet as taxonomy backbone. In Section 5 we outline the features used in the system and in section 6 the classifier.

## 5 FEATURES EXTRACTION

Every audio sample is converted to 22.05 KHz mono and then passed through a noise gate in order to determine its beginning and its end. After a frame-

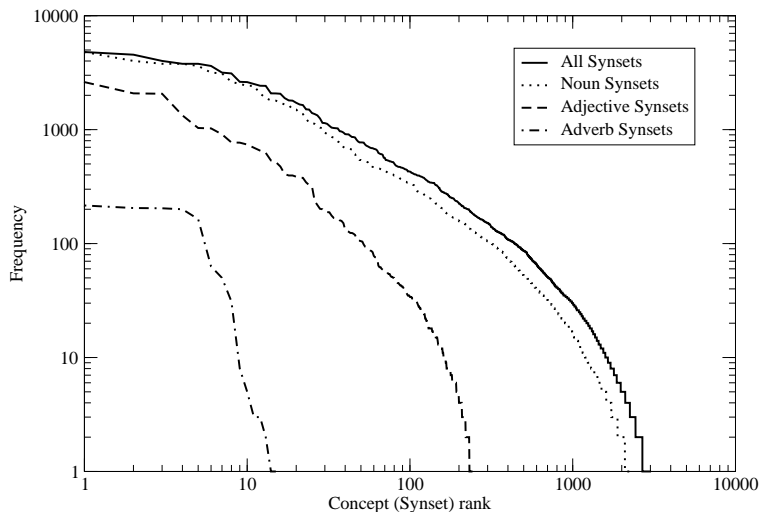


Figure 2: Number of sounds described per synset as a function of the synset rank. The frequency rank is plotted for the different parts of speech: noun, verb, adjective and adverbs.

by-frame analysis we extract features belonging to three different groups: a first group gathering spectral as well as temporal descriptors in the MPEG-7 standard; a second one built on Bark Bands perceptual division of the acoustic spectrum and which outputs the mean and variance of relative energies for each band; and, finally a third one, composed of Mel-Frequency Cepstral Coefficients and their corresponding variances (see Appendix and (Herrera, et al. 2002) for details).

## 6 NEAREST-NEIGHBOR CLASSIFIER

We use the  $k=1$  nearest neighbor decision rule (1-NN)(Jain, et al. 2000) for classification. The choice of a memory-based nearest neighbor classifier avoids the design and training of every possible class of sound (the order of several thousands). Another advantage of using a NN classifier is that it does not need to be redesigned, nor trained whenever a new class of sounds is subsequently added to the system. The NN classifier needs a database of labeled instances and a similarity distance to compare them. An unknown sample will borrow the metadata associated with the most similar registered sample. The similarity measure of the system is a normalized Manhattan distance of the above enumerated features:

$$d(x, y) = \sum_{k=1}^N \frac{|x_k - y_k|}{(\max_k - \min_k)}$$

	SN	TO	HH	CR	KI	RI
SN	150	1	2	2	1	20
TO	1	148	2	0	19	0
HH	5	7	153	0	1	4
CR	21	0	2	45	0	12
KI	1	17	0	0	182	0
RI	15	0	5	4	0	135

Table 1: Percussive instruments confusion matrix where SN:Snare, To:Tom, HH:Hihat, CR:Crash, KI:Kick, RI:Ride

Query Sound Caption	Nearest-neighbor Caption
Mini Cooper Door Closes Interior Persp.	Trabant Car Door Close
Waterfall Medium Constant	Extremely Heavy Rain Storm Short Loop
M-domestic Cat- Harsh Meow	A1v:Solo violin (looped)
Auto Pull Up Shut Off Oldsmobile	Ferrari - Hard Take Off Away - Fast
Animal-dog-snarl-growl-bark-vicious	Dinosaur Monster Growl Roar

Table 2: The classifier assigns the metadata of the sounds of the second column to the sounds of the first.

where  $x$  and  $y$  are the vectors of features,  $N$  the dimensionality of the feature space, and  $max_k$  and  $min_k$  the maximum and minimum values of the  $k$ th feature.

In some of our experiments, the standard deviation-normalized Euclidean distance does not perform well. Specially harmful is the normalization with standard deviation. Changing the normalization from the standard deviation to the difference between maximum and minimum boosted classification. For example the percussive instrument classification (see Section 7) raises from 64% to 82% correct identification. Changing the distance from Euclidean to Manhattan provided an extra 3% improvement..

## 7 EXPERIMENTAL RESULTS

The first experiment consisted in finding a best-match for all the sounds in the database. Table 2 shows some examples: on the left column the original caption of the sound and on the right the caption of the nearest neighbor. The caption on the right would be assigned to the query sound in an automatic annotation system.

As can be inferred from Table 2, it is not trivial to quantitatively evaluate the performance of the system. An intersection on the terms of the captions would not yield a reasonable evaluation metric. The WordNet based taxonomy can inform us that both “Trabant” and “Mini Cooper” are narrow terms for the concept “car, automobile”. Thus, the comparison of number of common synsets on both query and nearest-neighbor could be used as a better evaluation. As was shown in previous work (Cano, et al. 2004a), the intersection of synsets between query and best-match is 1.5 in average, while 50% of the times the best-match did not share a single common synset (see Figure 3). The intersection of source descriptions can be zero for very similar sounding sounds. The closest-

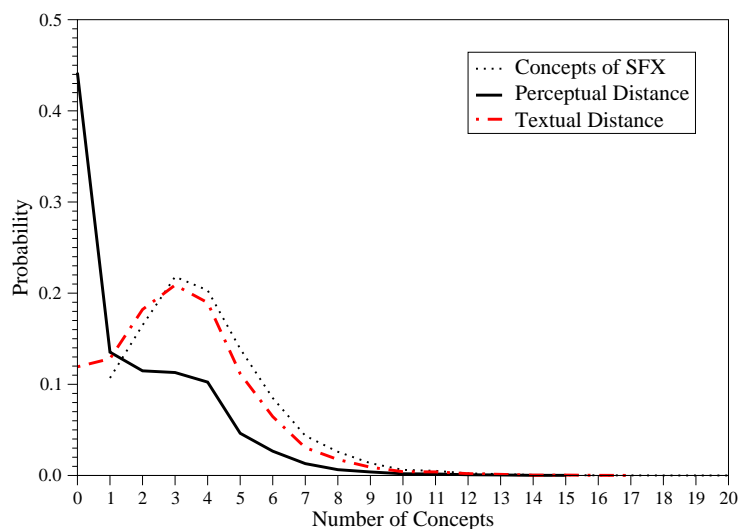


Figure 3: Probability distribution of correctly identified synsets. For each sound we count the intersection of concepts correctly predicted. The Concepts per sound shows the perfect score. The perceptual distance prediction plot indicates the prediction accuracy using the perceptual similarity distance. The textual distance line indicates the prediction using the textual captions and a cosine distance and it is shown for comparison.

match for a “paper bag” turns out to be a “eating toast”. These sounds are semantically different but perceptually similar. This situation is very common, *sound engineers* take advantage of the ambiguity and use “coconut half-shells” to create the sound of a “horse’s hoof-beats” (L.Mott 1990). This ambiguity is a disadvantage when designing and assessing perceptual similarity distances.

Another experiment is the prediction of synsets, that is, how well a particular concept, say “cat miaow”, will retrieve “miaow” sounds. The methodology is as follows. For each synset, we retrieve the sounds that have been labeled with that particular synset. For each sound its nearest-neighbor is calculated. We finally compute how many best-matching sounds are also labeled with that synset. From the total of 3028 synsets we restricted the experiment to the ones that had been attached to 10 or more sounds. This leaves us with 1645 synsets. Figure 4 displays the results. The top figure displays how often a synset retrieved sounds whose best-matches were also labeled with that synset. The bottom figure, on the other hand, shows the precision on the best 20 retrieved sounds. The ordering of synsets on the x-axis corresponds to their frequency rank as displayed in Figure 2. It is interesting to see that there is not a strong correlation between the synset frequency and the precision. On a random guess one would expect some synsets predicted much better only because they are very frequent.

	AF	AS	BF	BT	BA	BC	CE	DB	EC	FL	HO	OB	PI	SS	TT
AF	7	0	3	0	0	0	0	0	0	1	0	0	0	0	0
AS	0	18	0	0	0	1	0	0	0	0	0	0	0	0	0
BF	0	0	9	0	0	0	0	0	0	0	0	0	0	1	0
BT	0	0	0	9	0	0	0	0	0	0	0	0	0	0	1
BA	0	0	0	0	14	0	0	0	0	0	0	0	0	1	0
BC	0	0	0	1	0	10	1	1	0	0	0	1	0	0	0
CE	0	1	0	0	0	1	74	3	0	0	0	0	0	0	0
DB	0	0	0	0	0	0	0	72	0	0	0	0	0	0	0
EC	0	1	1	0	0	2	0	0	5	1	0	1	0	2	1
FL	1	2	0	3	0	1	0	0	0	11	0	4	0	0	0
HO	0	0	0	0	2	0	0	0	0	0	10	0	0	0	0
OB	0	1	0	0	0	0	2	0	0	0	1	7	0	0	1
PI	0	0	0	0	0	0	0	0	0	0	0	0	87	0	0
SS	0	0	0	0	0	1	0	0	0	0	0	0	0	24	0
TT	0	0	0	2	0	0	0	0	0	0	0	0	0	0	7

Table 3: Harmonic instruments confusion matrix where AF:AltoFlute, AS:AltoSax, BF:BassFlute, BT:BassTrombone, BA:Bassoon, BC:BbClarinet, CE:Cello, DB:DoubleBass, EC:EbClarinet, FL:Flute, HO:Horn, OB:Oboe, PI:Piano, SS:SopranoSax, TT:TenorTrombone.

In a second experiment we tested the general approach in reduced domain classification regime mode: percussive instruments, harmonic instruments and we achieve acceptable performances. The assumption is that there is a parallelism between semantic and perceptual taxonomies in musical instruments. The psychoacoustic studies of (Lakatos 2000) revealed groupings based on the similarities in the physical structure of instruments. We have therefore evaluated the similarity with classification on the musical instruments space, a subspace of the universe of sounds.

Table 3 depicts the confusion matrix of a 15 class harmonic instrument classification which corresponds with a 77.3% (261 audio files). In the 6 class percussive instrument classification an 85% Recognition (955 audio files) using 10 fold validation (see Table 1).

The last experiment is the robustness of the NN classification framework to audio distortions. The harmonic instruments samples of the experiments of Table 3 have been transcoded and resampled into WAV PCM format and Ogg format<sup>2</sup>. The results are depicted in Table 4. The percentages indicate the classification accuracy using different audio qualities. The columns are the audio qualities used as reference. The rows indicate the audio qualities used in the queries.

## 8 DISCUSSION

A major issue when building sound classification systems is the need of a taxonomy that organizes concepts and terms unambiguously. If the task is classifying any possible sound, the taxonomy design becomes a daunting task. We need a taxonomy or classification scheme that encodes the common sense knowledge

<sup>2</sup><http://www.vorbis.com>



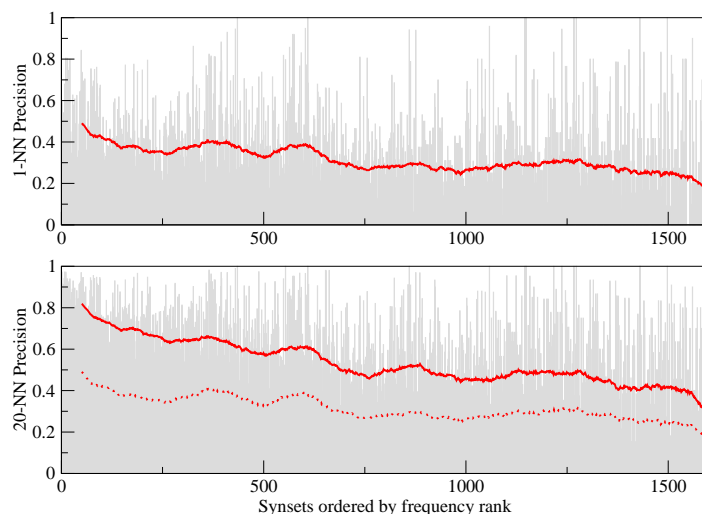


Figure 4: Synset precision using the 1-NN perceptual distance. The X axis corresponds to the synsets ordered by its frequency rank. The graph at the top shows the precision on the 1-NN. The bottom graph displays the precision on the 20 best retrieved sounds. The plots have been smoothed with an average filter. The dotted line of the bottom graph reproduces the precision on the 1-NN of the top graph.

of the world. WordNet can be used as a starting taxonomy. Normally, in identification a classifier is build to identify certain concepts: “cars”, “laughs”, “piano”. Sound samples are gathered and are tagged with those concepts and a classifier is trained to learn that concept. The number of concepts and its possible combinations in the real world makes this approach unfeasible, one would need to train tens of thousands of classifiers and new ones would have to be trained for new concepts. We have presented an alternative approach that uses an unambiguously labelled big audio database. The classifier uses nearest-neighbor rule and a database of sounds with WordNet as taxonomy backbone. As a results the list of possible sources is presented to the user: this sound could be a “paper bag” or “toast”+“eating”. Information from text or images can be used to disambiguate the possibilities.

We acknowledge that the use a single set of features and a single distance for all possible sound classes is rather primitive. However, as Figure 4 indicates, there is room for improvement. The NN rule can be combined with other classifiers: If the system returns that a particular sound could be a violin pizzicato or a guitar, we can then retrieve pizzicato violin and guitar sounds of the same pitch and train a classifier to decide which is more likely. Another example is “car approaches”, we can look for other “cars” and other “motor vehicle” “approaches” or “departs” to decide which is the right action. This same thinking

	Wav 44kHz	Ogg 44kHz	Ogg 11kHz
Wav 44kHz	91.5%	92.0%	75.0%
Wav 22kHz	86.4%	85.6%	82.0%
Wav 11kHz	71.8%	73.1%	89.3%
Ogg 44kHz	90.3%	91.5%	76.0%
Ogg 11kHz	74.0%	74.8%	91.5%

Table 4: Retrieval consistency on different distortions on the harmonic instruments classification. The columns indicate the reference audio quality and the rows the performance with the different distortions. Wav: PCM Microsoft WAV format, Ogg: Ogg Vorbis encoding, #kHz: Sampling rate

applies to adjective type of modifiers, something can be described as “loud”, “bright” or “fast”. The concept “fast” means something different if we talk of “footseps” or “typing”.

The system can be publicly accessed and tested through a web interface which allows users to upload sounds at <http://www.audioclas.org>.

## 9 Acknowledgments

We thank the staff from the Tape Gallery for all the support, discussion and feedback. This work is partially funded by the AUDIOCLAS Project E! 2668 Eureka. We thank the review and feedback from Fabien Gouyon.

## References

- K. Barnard, et al. (2003). ‘Matching Words and Pictures’. *Journal of Machine Learning Research* **3**:1107–1135.
- P. Cano, et al. (2004a). ‘Nearest-neighbor Generic Sound Classification with a WordNet-based taxonomy’. In *Proc. 116th AES Convention*, Berlin, Germany.
- P. Cano, et al. (2004b). ‘Sound Effects Taxonomy Management in Production Environments’. In *Proc. AES 25th Int. Conf.*, London, UK.
- M. Casey (2002). ‘Generalized Sound Classification and Similarity in MPEG-7’. *Organized Sound* **6**(2).
- S. Dubnov & A. Ben-Shalom (2003). ‘Review of ICA and HOS Methods for Retrieval of Natural Sounds and Sound Effects’. In *4th International Symposium on Independent Component Analysis and Blind Signal Separation*, Japan.
- B. Gygi (2001). *Factors in the identification of environmental sounds*. Ph.D. Thesis, Indiana University.
- P. Herrera, et al. (2003). ‘Automatic Classification of Musical Instrument Sounds’. *Journal of New Music Research* **32**(1).

Edges (Hz)	0	100	200	300	400	510	630	770	920	1080	1270	1480	1720
Centers (Hz)		50	150	250	350	450	570	700	840	1000	1170	1370	1600
Edges (Hz)	2000	2320	2700	3150	3700	4400	5300	6400	7700	9500	12000	15500	
Centers (Hz)	1850	2150	2500	2900	3400	4000	4800	5800	7000	8500	10500	13500	

Table 5: Bark band edges and centers.

- P. Herrera, et al. (2002). ‘Automatic Classification of Drum Sounds: A Comparison of Feature Selection Methods and Classification Techniques’. In C. Anagnostopoulou, M. Ferrand, & A. Smaill (eds.), *Music and Artificial Intelligence*. Springer.
- A. K. Jain, et al. (2000). ‘Statistical Pattern Recognition: A Review’. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(1):4–37.
- B. Kostek & A. Czyzewski (2001). ‘Representing Musical Instrument Sounds for Their Automatic Classification’. *J. Audio Eng. Soc.* **49**(9):768–785.
- S. Lakatos (2000). ‘A common perceptual space for harmonic and percussive timbres’. *Perception & Psychoacoustics* (62):1426–1439.
- R. L.Mott (1990). *Sound Effects: Radio, TV, and Film*. Focal Press.
- B. Logan (2000). ‘Mel frequency cepstral coefficients for music modeling’. In *Proc. of the ISMIR*, Plymouth, MA.
- K. D. Martin (1999). *Sound-Source Recognition: A Theory and Computational Model*. Ph.D. Thesis, M.I.T.
- G. A. Miller (November 1995). ‘WordNet: A Lexical Database for English’. *Communications of the ACM* pp. 39–45.
- G. Peeters & X. Rodet (2003). ‘Hierarchical Gaussian Tree with Inertia Ratio Maximization for the Classification of Large Musical Instruments Databases’. In *Proc. of the 6th Int. Conf. on Digital Audio Effects*, London.
- V. Peltonen, et al. (2002). ‘Computational Auditory Scene Recognition’. In *Proc. of ICASSP*, Florida, USA.
- M. Slaney (2002). ‘Mixture of Probability Experts for Audio Retrieval and Indexing’. In *IEEE International Conference on Multimedia and Expo*.
- E. Wold, et al. (1996). ‘Content-Based Classification, Search, and Retrieval of Audio’. *IEEE Multimedia* **3**(3):27–36.
- T. Zhang & C.-C. J. Kuo (1999). ‘Classification and Retrieval of Sound Effects in Audiovisual Data Management’. In *Proceedings of the 33rd Asilomar Conference on Signals, Systems and Computers*.

## 10 Appendix

### 10.1 Spectro-temporal descriptors

*Spectral Flatness* is the ratio between the geometrical mean and the arithmetical mean of the spectrum magnitude.

$$SFM = 10. \log \frac{(\prod_{k=1}^{N/2} S_p(e^{j \frac{2\pi k}{N}}))^{1/N/2}}{\frac{1}{N/2} \sum_{k=1}^{N/2} S_p(e^{j \frac{2\pi k}{N}})}$$

where  $S_p(e^{j \frac{2\pi k}{N}})$  is the spectral power density calculated on the basis of an  $N$ -point Fast Fourier Transform.

*Spectral Centroid* is a concept adapted from psychoacoustics and music cognition. It measures the average frequency, weighted by amplitude, of a spectrum. The standard formula for the (average) spectral centroid of a sound is:

$$c = \frac{\sum_j c_j}{J}$$

where  $c_j$  is the centroid for one spectral frame, and  $J$  is the number of frames for the sound. The (individual) centroid of a spectral frame is defined as the average frequency weighted by amplitudes, divided by the sum of the amplitudes.

$$c_j = \frac{\sum f_j a_j}{\sum a_j}$$

*Strong Peak* intends to reveal whether the spectrum presents a very pronounced peak.

*Spectral Kurtosis* is the spectrum 4th order central moment and measures whether the data are peaked or flat relative to a normal (gaussian) distribution.

$$kurtosis = \frac{\sum_{i=1}^N (Y_i - \bar{Y})^4}{(N-1)s^4}$$

where  $\bar{Y}$  is the sample mean,  $s$  is the sample standard deviation and  $N$  is the number of observations.

*Zero-Crossing Rate* (ZCR), is defined as the number of time-domain zero-crossings within a defined region of signal, divided by the number of samples of that region.

*Spectrum Zero-Crossing Rate* (SCR) gives an idea of the spectral density of peaks by computing ZCR at a frame level over the spectrum whose mean has previously been subtracted.

*Skewness* is the 3rd order central moment, it gives indication about the shape of the spectrum in the sense that asymmetrical spectra tend to have large Skewness values.

$$skewness = \frac{\sum_{i=1}^N (Y_i - \bar{Y})^3}{(N-1)s^3}$$

where  $\bar{Y}$  is the mean,  $s$  is the standard deviation, and  $N$  is the number of data points.

## 10.2 Bark-band energy

Bark-band energy are the energies after dividing the spectrum into the 24 Bark bands of frequencies depicted by Table 5. These bands are perception-related and have been chosen to enable systematic, instead of database-dependant, division of the spectrum. In order to cope with some low-frequency information that was found to be discriminative in a previous work (Herrera et al. 2002), the two lowest Bark bands have been splitted into two halves.

## 10.3 Mel-Frequency Cepstrum Coefficients

Mel-Frequency Cepstrum Coefficients (MFCCs) are widely used in speech recognition applications. They have been proved useful in music applications as well (Logan 2000). They are calculated as follows:

1. Divide signal into frames.
2. For each frame, obtain the amplitude spectrum.
3. Take the logarithm.
4. Convert to Mel spectrum.
5. Take the discrete cosine transform (DCT).

Step 4 calculates the log amplitude spectrum on the so-called Mel scale. The Mel transformation is based on human perception experiments. Step 5 takes the DCT of the Mel spectra. For speech, this approximates principal components analysis (PCA) which decorrelates the components of the feature vectors. Logan (Logan 2000) proved that this decorrelation applies to music signals as well. As they can be used as a compact representation of the spectral envelope, their variance was also recorded in order to keep some time-varying information. 13 MFCCs are computed frame by frame, and their means and variances are used as descriptors.

Serrà, J., Gómez, E., **Herrera, P.**, Serra, X. (2008). "Chroma binary similarity and local alignment applied to cover song identification". IEEE Transactions on Audio, Speech, and Language Processing, 16(6), pp. 1138-1151.

DOI: [10.1109/TASL.2008.924595](https://doi.org/10.1109/TASL.2008.924595)

ISSN: 1558-7916

Online ISSN: 1558-7924

# Chroma Binary Similarity and Local Alignment Applied to Cover Song Identification

Joan Serrà\*, Emilia Gómez, Perfecto Herrera and Xavier Serra

**Abstract**—We present a new technique for audio signal comparison based on tonal subsequence alignment and its application to detect cover versions (i.e., different performances of the same underlying musical piece). Cover song identification is a task whose popularity has increased in the Music Information Retrieval (MIR) community along in the past, as it provides a direct and objective way to evaluate music similarity algorithms. This article first presents a series of experiments carried out with two state-of-the-art methods for cover song identification. We have studied several components of these (such as chroma resolution and similarity, transposition, beat tracking or Dynamic Time Warping constraints), in order to discover which characteristics would be desirable for a competitive cover song identifier. After analyzing many cross-validated results, the importance of these characteristics is discussed, and the best-performing ones are finally applied to the newly proposed method. Multiple evaluations of this one confirm a large increase in identification accuracy when comparing it with alternative state-of-the-art approaches.

**Index Terms**—Music, Information retrieval, Acoustic signal analysis, Multidimensional sequences, Dynamic programming.

## I. INTRODUCTION

IN THE present times, any music listener may have thousands of songs stored in a hard disk or in a portable MP3 player. Furthermore, on-line digital music stores own large music collections, ranging from thousands to millions of tracks. Additionally, the ‘unit’ of music transactions has changed from the entire album to the song. Thus, users or stores are faced to search through vast music databases at the song level. In this context, finding a musical piece that fits one’s needs or expectancies may be problematic. Therefore, it becomes necessary to organize them according to some sense of similarity. It is at this point where determining if two musical pieces share the same melodic or tonal progression becomes interesting and useful. To address this issue, from a research perspective, a good starting point seems to be the identification of cover songs (or versions), where the relationship between them can be qualitatively defined, objectively measured, and is context-independent. In addition, from the users perspective, finding all versions of a particular song can be valuable and fun.

It is important to mention that the concept of music similarity, and more concretely, finding cover songs in a database, has a direct implication to musical rights management and

licenses. Also, learning about music itself, discovering the musical essence of a song, and other many topics related with music perception and cognition are partially pursued by this research. Furthermore, the techniques presented here can be exploited for general audio signal comparison, where cover/version identification is just an application among other possible ones.

The expressions *cover song* and *version* may have different and somehow fuzzy connotations. A *version* is intended to be what every performer does by playing precomposed music, while the term *cover song* comes from a very different tradition in pop music, where a piece is composed for a single performer or group. Cover songs were, originally, part of a strategy to introduce ‘hits’ that had achieved significant commercial success from other sections of the record-buying public, without remunerating any money to the original artist or label. Nowadays, the term has nearly lost these purely economical connotations. Musicians can play covers as a homage or a tribute to the original performer, composer or band. Sometimes, new versions are made for translating songs to other languages, for adapting them to a particular country/region tastes, for contemporising familiar or very old songs, or for introducing new artists. In addition, cover songs represent the opportunity to perform a radically different interpretation of a musical piece.

Today, and perhaps not being the proper way to name it, a cover song can mean any new version, performance, rendition or recording of a previously recorded track [1]. Therefore, we can find several musical dimensions that might change between two covers of the same song. These can be related to timbre (different instruments, configurations or recording procedures), tempo (global tempo and tempo fluctuations), rhythm (e.g., different drum section, meter, swinging pattern or syncopation), song structure (eliminating introductions, adding solo sections, choruses, codas, etc.), main key (transposition to another tonality), harmonization (adding or deleting chords, substituting them by related ones, adding tensions, ...) and lyrics (e.g., different languages or words).

A robust mid-level characteristic that is largely preserved under the mentioned musical variations is a tonal sequence (or a harmonic progression [2]). Tonality is ubiquitous and most listeners, either musically trained or not, can identify the most stable pitch while listening to tonal music. Furthermore, this process is continuous and remains active throughout the sequential listening experience [3], [4]. From the point of view of the Music Information Retrieval (MIR) field, clear insights about the importance of temporal and tonal features in a music similarity task have been evidenced [5], [6], [7].

Tonal sequences can be understood as series of different

Manuscript received November 30, 2007; revised February 21, 2008; accepted April 5, 2008. This research has been partially funded by the EU-IP project PHAROS IST-2006-045035: <http://www.pharos-audiovisual-search.eu>  
J. Serrà, E. Gómez, P. Herrera and X. Serra are with the Music Technology Group, Universitat Pompeu Fabra, Ocata 1 (3rd floor), 08003 Barcelona, Spain, phone: +34-93-542-2864, fax: +34-93-542-2202, e-mail: {jserra,egomez,pherrera,xserra}@iua.upf.edu

note combinations played sequentially. These notes can be unique for each time slot (a melody) or can be played jointly with others (chord or harmonic progressions). Systems for cover song identification usually exploit these aspects and attempt to be robust against changes in other musical facets. In general, they either try to extract the predominant melody [8], [9], a chord progression [10], [11], or a chroma sequence [12], [13], [14], [15], [16]. Some methods do not take into account (at least explicitly) key transposition between songs [13], [14], but the usual strategy is to normalize these descriptor sequences in respect to the key. This is usually done by means of a key profile extraction algorithm [9], [10], [15], or by considering all possible musical transpositions [12], [8], [11], [16]. Then, for obtaining a similarity measure, descriptor sequences are usually compared by means of Dynamic Time Warping (DTW) [8], [10], [15], an edit-distance variant [7], [11], string matching [12], Locality Sensitive Hashing (LSH) [14], or a simple correlation function or a cosine angle [9], [13], [16]. In addition, a beat tracking method might be used [9], [12], [16], or a song summarization or chorus extraction technique might be considered [9], [15].

Techniques for predominant melody extraction have been extensively researched in the MIR community [17], [18], [19], as well as key/chord identification engines [20], [21]. Also, chroma-based features have become very popular [22], [23], [24], [25], with applications in various domains such as pattern discovery [26], audio thumbnailing and chorus detection [27], [28], or audio alignment [5], [29].

Regarding alignment procedures and sequence similarity measures, DTW [30] is a well known technique used in speech recognition for aligning two sequences which may vary in time or speed and for measuring similarity between them. Also, several edit-distance variants [31] are widely used in very different disciplines such as text retrieval, DNA or protein sequence alignment [32], or MIR itself [33], [34]. If we use audio shingles (i.e., high-dimensional feature vectors concatenations) to represent different portions of a song sequence, LSH solves fast approximate nearest neighbor search in high dimensions [35].

One of the main goals of this article is to present a study of several factors involved in the computation of alignments of musical pieces and similarity of (cover) songs. To do this, the impact of a set of factors in state-of-the-art cover song identification systems is measured. We experiment with different resolution of chroma features, with different local cost functions (or distances) between chroma features, with the effect of using different musical transposition methods, and with the use of a beat tracking algorithm to obtain a tempo-independent chroma sequence representation. In addition, as DTW is a well known and extensively employed technique, we test two underexplored variants of it: DTW with global and local constraints. All these experiments are oriented to elucidate the characteristics that a competitive cover song identification system should have. We then apply this knowledge to a newly proposed method, which uses sequences of feature vectors describing tonality (in our case Harmonic Pitch Class Profiles [25], from now on HPCP), but it presents relevant differences in two important aspects: we use a novel binary

similarity function between chroma features, and we develop a new local alignment algorithm for assessing resemblance between sequences.

The rest of the paper is organized as follows. First, in section II, we explain our test framework. We describe the methods used to evaluate several relevant parameters of a cover song identification system (chroma resolution and similarity, key transposition, beat tracking and DTW constraints), and the descriptors employed across all these experiments. We also introduce the database and the evaluation measures that are employed along this study. Then, in section III, we sequentially present all the evaluated parameters and the obtained results. In section IV, we propose a new method for assessing the similarity between cover songs. This is based on the conclusions obtained through our experiments (summarized in section III-F) and on two main aspects: a new chroma similarity measure and a novel dynamic programming local alignment algorithm. Finally, a short conclusions section closes the study.

## II. EXPERIMENTAL FRAMEWORK

### A. Tonality descriptors

All the implemented methods use the same feature set: sequences of *Harmonic Pitch Class Profiles* (HPCP) [25]. The HPCP is an enhanced pitch class distribution (or chroma) feature, computed in a frame-by-frame basis only using the local maxima of the spectrum within a certain frequency band. Chroma features are widely used in the literature and proven to work quite well for the task at hand [13], [15], [16]. In general, chroma features should be robust to noise (e.g., ambient noise or percussive sounds), independent of timbre and played instruments (so that the same piece played with different instruments has the same tonal description), and independent of loudness and dynamics. These are some of the qualities that might make them lead to better results for cover song identification when comparing them, for instance, with MFCCs [7], [14].

In addition to using the local maxima of the spectrum within a certain frequency band, HPCPs are tuning independent (so that the reference frequency can be different from the standard A 440 Hz), and consider the presence of harmonic frequencies. The result of HPCP computation is a 12, 24 or 36-bin (depending on the desired resolution) octave-independent histogram representing the relative intensity of each 1, 1/2 or 1/3 of the 12 semitones of the equal tempered scale. A schema of the extraction process and a plot of the resulting HPCP sequence are shown in figures 1 and 2.

We start by cutting the song into short overlapping and windowed frames. For that, we use a Blackman-Harris (62 dB) window of 93 ms length with a 50% frame overlapping. We perform a spectral analysis using the Discrete Fourier Transform (DFT), and the spectrum is whitened by normalizing the amplitude values with respect to the spectral envelope. From the obtained spectrum, we compute a set of local maxima or peaks and we select the ones with frequency values  $f_i \in (40, 5000)$  Hz. The selected spectral peaks are summarized in an octave-independent histogram according to a reference frequency (around 440 Hz). This reference frequency is estimated by



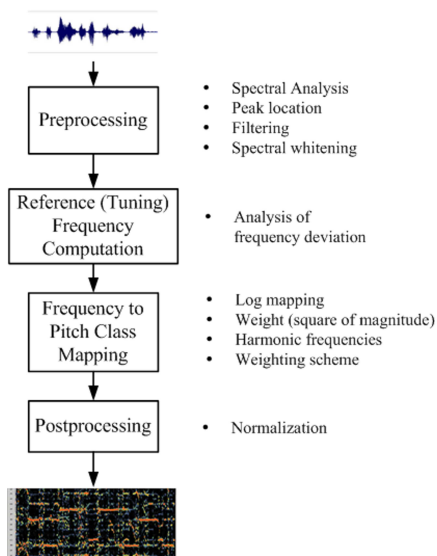


Fig. 1. General HPCP feature extraction block diagram. Audio (top) is converted to a sequence of HPCP vectors (bottom) that evolves with time.

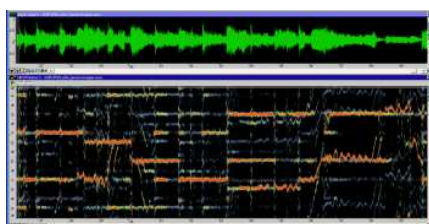


Fig. 2. Example of a high-resolution HPCP sequence (bottom panel) corresponding to an excerpt of the song “Imagine” by John Lennon (top panel). In the HPCP sequence, time (in frames) is represented in the horizontal axis and chroma bins are plotted in the vertical axis.

analyzing the deviations of the spectral peaks with respect to an equal-tempered chromatic scale. A global estimate of this reference frequency is employed for all the analyzed frames.

Instead of contributing to a single HPCP bin, each peak frequency  $f_i$  contributes to the HPCP bin(s) that are contained in a certain window around its frequency value. The peak contribution  $i$  is weighted using a  $\cos^2$  function around the bin frequency. The length of the weighting window  $l$  have been empirically set to  $\frac{4}{3}$  semitones. This weighting procedure minimizes the estimation errors that we find when there are tuning differences and inharmonicity present in the spectrum, which could induce errors when mapping frequency values into HPCP bins.

In addition, in order to make harmonics contribute to the pitch class of its fundamental frequency, we also introduce an additional weighting procedure: each peak frequency  $f_i$  has a contribution to its  $n$ Harmonics = 8 sub-harmonics. We make this contribution decrease along frequency using an

exponential function.

The HPCP extraction procedure employed here is the same that has been used in [15], [36], [37], [25], and the parameters mentioned in this paragraph have been proven to work well for key estimation and chord extraction in the previously cited references.

An exhaustive comparison between ‘standard’ chroma features and HPCPs is presented in [25] and [38]. In [25], a comparison of different implementations of chroma features (Constant-Q profiles [39], Pitch Class Profiles (PCP) [20], chromagrams [21] and HPCP) with MIDI-based Muse Data [40] is provided. The correlation of HPCP with Muse Data was higher than 0.9 for all the analyzed pieces (48 Fugues of Bach’s WTC) and HPCPs outperformed the Constant-Q profiles, chromagrams and PCPs. We also compared the use of different HPCP parameters, arriving to optimal results with the ones used in the present work. In [38], the efficiency of different sets of tonal descriptors for music structural discovery was studied. Herein, the use of three different pitch-class distribution features (i.e., Constant-Q Profile, PCP and HPCP) was explored to perform structural analysis of a piece of music audio. A database of 56 audio files (songs by The Beatles) was used for evaluation. The experimental results showed that HPCP were performing best, yielding an average of 82% of accuracy in identifying structural boundaries in music audio signals.

## B. Studied methods

We now describe two methods that have served us to test several important parameters of a cover song identification system, as a baseline for further improvements [16], [25]. We have chosen them because they represent in many ways the state-of-the-art. Their main features are the use of global alignment techniques and common feature dissimilarity measures. In subsequent sections, we differentiate these two methods by its alignment procedure (cross-correlation or Dynamic Time Warping), but other procedures are characteristic for each one (such as audio features, dissimilarity measure between feature vectors, etc.).

1) *Cross-correlation approach*: A quite straightforward approach is presented in [16]. This method finds cover versions by cross-correlating chroma vector sequences (representing the whole song) averaged beat-by-beat. It seems to be a good starting point since it was found to be superior to other methods presented to MIREX 2006 evaluation contest<sup>1</sup>. We worked with a similar version of the forementioned system. We re-implemented the algorithm proposed by the authors<sup>2</sup> in order to consider the same chroma features for all the methods (HPCPs) and to ease the introduction of new functionalities and improvements. We now describe the followed steps.

First of all, HPCP features are computed. Each frame vector is normalized by dividing it by its maximum amplitude, as shown in figure 1. In addition, beat timestamps are computed

<sup>1</sup>See the complete results at [http://www.music-ir.org/mirex/2006/index.php/Audio\\_Cover\\_Song](http://www.music-ir.org/mirex/2006/index.php/Audio_Cover_Song) (Accessed 28 Jan. 2008)

<sup>2</sup><http://labrosa.ee.columbia.edu/projects/coversongs> (Accessed 28 Jan. 2008)

with an algorithm adapted from [41], [42] using the *aubio* library<sup>3</sup>.

The next step is to average the frame-based HPCP vectors contained in between each two beat timestamps. With this, we obtain a tempo-independent HPCP sequence. In order to account for key changes, the two compared HPCP sequences are usually transposed to the same key by means of a key extraction algorithm or an alternative approach (see section III-C). Another option is the one proposed in [16], where the sequence similarity measure is computed for all possible transpositions and the maximum value is then chosen.

In this approach, sequence similarity is obtained through cross-correlation. That is, we calculate a simple cross-correlation between each two tempo-independent HPCP sequences for each song being compared (with possibly different lengths). The cross-correlation values are further normalized by the length of the shorter segment, so that the measure is bounded between zero and one. Note that a local distance measure between HPCPs must be used. The most usual thing is to use an euclidean-based distance, but other measures can be tried (see section III-B).

In [16], the authors found that genuine matches were indicated not only by cross-correlations of large magnitudes, but that these large values occurred in narrow local maxima in the cross correlations that fell off rapidly as the relative alignment changed from its best value. So, to maximize these local maxima, cross-correlation was high-pass filtered. Finally, the final measure representing the dissimilarity between two songs is obtained with the reciprocal of the maximum peak value of this high-pass filtered cross-correlation.

2) *Dynamic Time Warping approach*: Another approach for detecting cover songs was implemented, reflecting the most used alignment technique in the literature: Dynamic Time Warping (DTW). The followed method has a very high resemblance with the one presented in [25].

We proceed by extracting HPCP features in the same way as the previous approach (section II-B1). Here, we do not use any beat tracking method because DTW is specially designed for dealing with tempo variations (see section III-D). For speeding up calculations, a usual strategy is to average each  $k$  consecutive descriptors vectors (frames). We call this value ( $k$ ) the *averaging factor*. Here, each HPCP feature vector is also normalized by its maximum value. We deal with key invariance just in the same way than the previous approach (section II-B1) and transpose the HPCP sequences representing the two songs' tonal progressions to a common key.

To align these two sequences (which can have different lengths  $n$  and  $m$ ), we use the DTW algorithm [30]. It basically operates by recursively computing an  $n \times m$  cumulative distance matrix by using the value of a local cost function. This local cost function is usually set to be any euclidean-based distance, though in [15], [25] the correlation between the two HPCP vectors is used to define the dissimilarity measure (see section III-B). With DTW, we obtain the total alignment cost between two HPCP sequences in matrix element  $(n, m)$ .

We can also obtain an alignment path whose length acts as a normalization factor.

### C. Evaluation methodology

To test the effectiveness of the implemented systems under different parameter configurations, we compiled a music collection comprising 2053 commercial songs distributed in different musical genres. Within these songs, there were 451 original pieces (we call them *canonical versions*) and 1462 covers. Songs were obtained from personal music collections. The average number of covers per song was 4.24, ranging from 2 (the original song plus 1 cover) to 20 (the original song plus 19 covers). There were also 140 'confusing songs' from the same genres and artists as the original ones that were not associated to any cover group. A special emphasis was put in the variety of styles and the employed genres for each cover set. A complete list of the music collection can be found in our web page<sup>4</sup>.

Due to the high computational cost of the implemented cover song identification algorithms, we have restricted the music collection for preliminary experiments. We simultaneously employed two non-overlapping smaller subsets of the whole song database, intended to be as representative as possible of the entire corpus. We provide some statistics in table I.

TABLE I  
SONG COMPILATIONS USED. DB75, DB330 AND DB2053 CORRESPOND TO THE NAMES WE GIVE TO THE DIFFERENT DATABASES. '\*' DENOTES AVERAGE NUMBER OF COVERS PER GROUP. IN DB75 AND DB330 THERE WERE NO 'CONFUSING SONGS'

	DB75	DB330	DB2053
Total number of songs	75	330	2053
Number of cover sets	15	30	451
Covers per set	5	11	4.24*

We queried all the covers and canonical versions and obtained a distance matrix whose dimensions depended on the number of songs. This data was further processed in order to obtain several evaluation measures. Here, we mainly show the results corresponding to standard F-measure and average Recall ( $R_x$ ) [43]. This last measure was computed as the mean percentage of identified covers within the first  $x$  answers. All experiments were evaluated with these measures, and, most of the time, other alternative metrics were highly correlated with the previous ones. A qualitative assessment of valid evaluation measures for this cover song system was presented in [44].

## III. EXPERIMENTS

The next subsections describe the tests carried out to evaluate the impact of several system parameters and procedures in both methods explained in section II-B. Our hypothesis was that these had a strong influence in final identification accuracy and shouldn't be blindly assigned. To our knowledge, this is one of the first systematic study of this kind that has been made until now (with, perhaps, the exception of [11], where the author evaluated the influence of key shifting, cost gap

<sup>3</sup><http://aubio.org> (Accessed 28 Jan. 2008)

<sup>4</sup><http://mtg.upf.edu/~jserra/files/coverdatabase.csv.tar.gz>

insertions and character swaps in a string alignment method used for cover song identification, in addition to the use of a beat-synchronous set).

In our experiments, we aimed at measuring, on a state-of-the-art cover song identification system, the impact of the following factors [45]: (a) the resolution of the chroma features, (b) the local cost function (or distance) between chroma features, (c) the effect of using different key transposition methods, and (d) the use of a beat tracking algorithm to obtain a tempo-independent chroma sequence representation. In addition, as DTW is a well known and extensively employed technique, we wanted to (e) test two underexplored variants of it: DTW with global and local constraints. A wrap-up discussion on these factors is provided in section III-F. Finally, we want to highlight that through all experiments reported in this section, all combinations of parameters cited in each subsection were studied. We report average performance results for each subsection given that all parameter combinations resulted in similar behaviours. Different behaviours are properly highlighted through the text, if any.

#### A. Effect of chroma resolution

Usually, chroma features are represented in a 12-bin histogram, each bin corresponding to 1 of the 12 semitones of the equal-tempered scale. But higher resolutions can be used to get a finer pitch class representation. Other commonly used resolutions are 24 and 36 bins [25] (corresponding to 1/2 or 1/3 of a semitone). We tested these three values in our experiments. The resolution parameter was changed in the HPCP extraction method of the approaches explained in section II-B.

The average identification accuracy across experiments with two different chroma similarity measures (section III-B) and two key transposition methods (section III-C) are shown in table II. In all the experiments, and independently of the HPCP distance used and the transposition made, the greater the HPCP resolution, the better the accuracy we got (F-measure more than 12% better).

TABLE II  
F-MEASURE AND AVERAGE RECALL WITHIN THE FIRST FOUR RETRIEVED SONGS FOR DIFFERENT HPCP RESOLUTIONS. AVERAGE OF DIFFERENT CROSS-CORRELATION APPROACH VARIANTS EVALUATED WITH DB75

Resolution	F-measure	R <sub>4</sub>
12 bins	0.495	0.429
24 bins	0.511	0.435
36 bins	0.558	0.489

#### B. Effect of chroma similarity measures

In order to test the importance of the used HPCP distance measure, we evaluated two similarity measures: cosine similarity and the correlation between feature vectors. These two measures were chosen because they are commonly used in the literature. Correlation has been used in [15], [25], and is inspired on the cognitive aspects of pitch processing in humans [46]. Furthermore, for key extraction, it was found to

work better than the simple euclidean distance between HPCP vectors [25].

Tests were made with the methods exposed in section II-B and the two measures cited above. The results are shown in table III. We observe that the employed HPCP distance plays a very important role. This aspect of the system can yield to more than a 13% accuracy improvement for some tests [45]. In all trials made with different resolutions and ways of transposing songs, correlation between HPCPs was found to be a better similarity measure than cosine distance<sup>5</sup>. The former gives a mean F-measure improvement, among the tested variants, of approximately 6%.

TABLE III  
F-MEASURE AND AVERAGE RECALL WITHIN THE FIRST FOUR RETRIEVED SONGS FOR COSINE DISTANCE ( $d_{COS}$ ) AND CORRELATION DISTANCE ( $d_{CORR}$ ). AVERAGE OF DIFFERENT CROSS-CORRELATION APPROACH VARIANTS EVALUATED WITH DB75

Distance used	F-measure	R <sub>4</sub>
$d_{COS}$	0.504	0.436
$d_{CORR}$	0.537	0.461

#### C. Effect of key transposition

In order to account for songs played in a different key than the original one, we calculated a global HPCP vector and we transposed (circularly-shifted) one HPCP sequence to the other's tonality. This procedure was introduced in both methods described in section II-B. A global HPCP vector was computed by averaging all HPCPs in a sequence, and it was normalized by its maximum value as all HPCPs. With the global HPCPs of two songs ( $\vec{h}_A$  and  $\vec{h}_B$ ), we computed what we call the *Optimal Transposition Index* (from now on OTI), which represents the number of bins that an HPCP needs to be circularly shifted to have maximal resemblance to the other:

$$OTI(\vec{h}_A, \vec{h}_B) = \operatorname{argmax}_{0 \leq id \leq N_H - 1} \{ \vec{h}_A \cdot \operatorname{circshift}_R(\vec{h}_B, id) \} \quad (1)$$

where  $\cdot$  indicates a dot product,  $N_H$  is the HPCP size considered, and  $\operatorname{circshift}_R(\vec{h}, id)$  is a function that rotates a vector ( $\vec{h}$ )  $id$  positions to the right. A circular shift of one position is a permutation of the entries in a vector where the last component becomes the first one and all the other components are shifted. Then, to transpose one song, for each HPCP vector  $i$  in the whole sequence we compute:

$$\vec{h}_{A,i}^{Tr} = \operatorname{circshift}_R(\vec{h}_{A,i}, OTI) \quad (2)$$

where superscript  $Tr$  denotes musical HPCP transposition.

In order to evaluate the goodness of this new procedure for transposing both songs to a common key, an alternative way of computing a transposed HPCP sequence was introduced. This consisted on calculating the main tonality for each piece using a key estimation algorithm [25]. This algorithm is a state-of-the-art approach with an accuracy of 75% for real

<sup>5</sup><http://mtg.upf.edu/~jserra/chromabinsimappendix.html>

audio pieces [36], and scored among the first classified algorithms in the MIREX 2005 contest<sup>6</sup> with an accuracy of 86% with synthesized MIDI files. With this alternative procedure, once the main tonality was estimated, the whole song was transposed according to this estimated key. A possibly better way of dealing with key changes would be to calculate the similarity measures for all possible transpositions and then take the maximum [16]. We have not tested this procedure since for high HPCP resolutions it becomes computationally expensive.

OTI and key transposition methods were compared across several HPCP resolutions (section III-A) and two different HPCP distance measures (section III-B). The averaged identification accuracy is shown in table IV. It can be clearly seen that a key estimation algorithm has a detrimental effect to overall results (F-measure 17% worse). This was also independent of the number of bins and the HPCP distance used<sup>7</sup>. We have evaluated dependence of the number of HPCP bins, and HPCP distance, and we have found that they had similar behavior. Therefore, it seems appropriate to transpose the songs according to the OTI of the global HPCP vectors. Apart from testing the appropriateness of our transposition method, we were also pursuing the impact that different transposition methods could have, which we see is quite important in table IV.

TABLE IV  
F-MEASURE AND AVERAGE RECALL WITHIN THE FIRST FOUR RETRIEVED SONGS FOR GLOBALHPCP + OTI TRANSPOSITION METHOD AND BY USING A KEY ESTIMATION ALGORITHM. AVERAGE OF DIFFERENT CROSS-CORRELATION APPROACH VARIANTS EVALUATED WITH DB75

Method	F-measure	R <sub>4</sub>
GlobalHPCP + OTI	0.569	0.500
Key finding algorithm	0.474	0.400

#### D. Effect of beat tracking and averaging factors

In the cross-correlation approach (section II-B1), HPCP vectors were averaged beat-by-beat. With the DTW approach of section II-B2, we expected DTW being able to cope with tempo variations. To demonstrate this, we performed some tests with DTW. In these, several *averaging factors* were also tried.

Experiments were done with 5 different DTW algorithms (see section III-E). In these and subsequent experiments HPCP resolution was set to 36, correlation was used to assess the similarity between HPCP vectors and we employed OTI-based transposition. Results shown in table V are the average identification accuracy values obtained across these different implementations. We have to note that taking the arithmetic mean of the respective evaluation measures masks the concrete behaviour of them along different averaging factors (information regarding the effect of different averaging factors upon considered constraints can be found in subsequent section III-E). Nevertheless, for all the tested variants, better

accuracies were reached with averaging HPCPs in a frame basis, than using beat-by-beat averaging. A similar result using the Needleman-Wunsch-Sellers algorithm [47] reported in [11] supports our findings.

TABLE V  
F-MEASURE AND AVERAGE RECALL WITHIN THE FIRST FOUR RETRIEVED SONGS FOR DIFFERENT *averaging factors* (INCLUDING BEAT AVERAGING). CORRESPONDING TIME FACTOR IS EXPRESSED IN THE SECOND COLUMN. AVERAGE OF DIFFERENT DTW APPROACH VARIANTS EVALUATED WITH DB75

Averaging factor (frame count)	Averaging length (seconds)	F-measure	R <sub>4</sub>
Beat	variable	0.469	0.417
5	0.232	0.470	0.419
10	0.464	0.494	0.448
15	0.696	0.511	0.465
20	0.929	0.514	0.463
25	1.161	0.512	0.466
30	1.393	0.510	0.461
40	1.856	0.487	0.434

#### E. Effect of DTW global and local constraints

We can apply different constraints to a DTW algorithm in order to decrease the number of paths considered during the matching process. These constraints are desirable for two main purposes: to reduce computational costs and to prevent ‘pathological’ warpings. ‘Pathological’ warpings are considered the ones that, in an alignment, assign several multiple values of a sequence to just one value of the other sequence. This is easily seen as a straight line in the DTW matrix (an example is shown in the first plot of figure 3).

To test the effect of these constraints we implemented 5 variants of a DTW algorithm: the one mentioned in section II-B2, two globally constrained DTW algorithms, and two locally constrained ones:

- Simple DTW: This implementation corresponds to the standard definition of DTW, where no constraints are applied [30].
- Globally constrained DTW: Two implementations were tried. One corresponds to Sakoe-Chiba constraints [48] and the other one to the Itakura parallelogram [49]. With these global constraints, elements far from the diagonal of the  $n \times m$  DTW matrix are not considered (see figure 4). A commonly used value for that in many speech recognition tasks is 20% [30].
- Locally constrained DTW: To further specify the optimal path, some local constraints can be applied in order to guarantee that excessive time scale compression or expansion is avoided. We specified two local constraints that were found to work in a plausible way with speech recognition [50]. From this reference, *Type 1* and *Type 2* constraints were chosen (we denote them *MyersT1* and *MyersT2* respectively). For both, the recursive relation of DTW is changed in such a way that in element  $(i, j)$  of a DTW cumulative distance matrix, we only pay attention to warpings  $(i-1, j-1)$  (no tempo deviation),  $(i-2, j-1)$  ( $2x$  tempo deviation) and  $(i-1, j-2)$  ( $0.5x$  tempo deviation). So, we allow maximal deviations of the double

<sup>6</sup>[http://www.music-ir.org/mirex/2005/index.php/Audio\\_and\\_Symbolic\\_Key\\_Finding](http://www.music-ir.org/mirex/2005/index.php/Audio_and_Symbolic_Key_Finding) (Accessed 29 Jan. 2008)  
<sup>7</sup><http://mtg.upf.edu/~jserra/chromabinsimappendix.html>

or half the tempo. This seems reasonable for us since, for instance, if the original song is at 120 B.P.M., a cover may not be at less than 60 B.P.M. or more than 240 B.P.M. The difference between *MyersT1* and *MyersT2* constraints relies in the way we weight this warpings: considering intermediate distances for the former, and double-weighting the distance between elements  $i$  and  $j$  for the latter [50].

These three implementations were evaluated across different averaging factors (see section III-D) and the means of the F-measure and average recall within the 4 first answered items ( $R_4$ ) were taken. Results can be seen in table VI. In general, better accuracies are achieved with local constraints, whereas global constraints yielded the worst results.

TABLE VI  
F-MEASURE AND AVERAGE RECALL WITHIN THE FIRST FOUR RETRIEVED SONGS FOR DIFFERENT DTW ALGORITHMS IMPLEMENTING GLOBAL AND LOCAL CONSTRAINTS. EVALUATION WAS DONE WITH DB75

Alg. name	Constr. type	F-measure	$R_4$
Sakoe-Chiba	Global	0.321	0.283
Itakura	Global	0.344	0.304
Simple DTW	No constr.	0.600	0.541
MyersT2	Local	0.608	0.552
MyersT1	Local	0.624	0.570

There is one important fact about local constraints that needs to be remarked and that can be appreciated in table VII. In general (except for the locally constrained methods), as the framelength decreases, it can be seen that identification accuracy does so. This is due to the fact that lower framelengths introduce the creation of ‘pathological’ warping paths (straight lines in the DTW matrix) that do not correspond to the true alignment (a straight line indicates several points of one sequence aligned just to one point of the other, left picture in figure 3). This makes the path length to increase, and since we normalize the final result by this value to yield sequence length independence, the final distance value decreases. Then, false positives are introduced in the final outcomes of the algorithm. Figure 3 shows the same part for matrices obtained after a simple and a locally constrained DTW approach. Local constraints prevent DTW from these undesired warpings. If there is a single horizontal or vertical step in the warping path, they force them to be the opposite way in next recurrent step. This is why the accuracy of locally constrained methods keeps increasing while lowering the averaging factor.

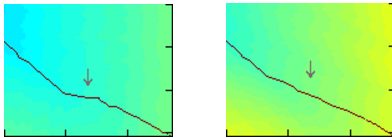


Fig. 3. Parts of the matrix obtained with a simple (left) and locally constrained (MyersT1, right) DTW approach for the same two songs. On the left we can observe some ‘pathological’ warpings, while on the right, these have disappeared.

Also in table VII, we observe that the identification accuracy for globally constrained methods is significantly lower than for

the other ones. This is due to the fact that, by using these global constraints, we restrict the paths to be around the DTW matrix main diagonal. To understand the effect of that, as an example, we consider a song composed by two parts that are the same ( $S_1 = AA$ ) and another song (a cover) with nearly half the tempo ( $S_2 = A'$ ) and composed by only one of these parts ( $S_2 = A'$ ). The plots in figure 4 graphically explain this idea. The first one (left) was generated using a method with no constraints. We observe that the best path (straight diagonal red line) goes from  $(1, 1)$  to more or less  $(20, 10)$  (horizontal axis lower-half part). This is logical since  $S_2$  (vertical axis) is a half-tempo kind-of repetition of one part of  $S_1$  (horizontal axis). The middle plot corresponds to the same matrix with Sakoe-Chiba constraints. We observe that the ‘optimal’ path we could trace with the first plot has been broken by the effect of the global constraints. A similar situation occurs with Itakura constraints (right plot).

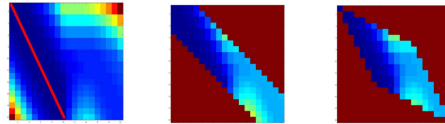


Fig. 4. Examples of an unconstrained DTW matrix (left), and Sakoe-Chiba (center) and Itakura (right) global constraints for  $S_1$  ( $x$ -axis) and  $S_2$  ( $y$ -axis). As this is an intuitive example, coordinate units in the horizontal and vertical axes are arbitrary.

## F. Discussion

In previous subsections we have studied the influence of several aspects in two state-of-the-art methods for cover song identification. All the analyzed features proved to have a direct (and sometimes dramatic) impact in the final identification accuracy. We are now able to summarize some of the key aspects that should be considered when identifying cover songs. These aspects have been considered as a basis to design our approach, which will be presented in the following sections.

1) *Audio features*: The different musical changes involved in cover songs, as discussed in section I, give us clear insights on which features to use. As chroma features have been evidenced to work quite well for this task [13], [15], [16] and proven to be better than timbre oriented descriptors as MFCC [7], [14], our approaches are based on HPCPs, given their usefulness for other tasks (e.g., key estimation) and their correspondence to pitch class distributions (see [25], [38] for a comparison with alternative approaches).

In section III-A, we have shown that HPCP resolution is important with both cosine and correlation distances. We have tested 12, 24, and 36-bin HPCPs with different variants of the methods presented in section II-B, and the results suggest that accuracy increases as the resolution does so. On the other hand, increasing resolution also increases computational costs, so that higher resolution is not considered. In addition, 36 seems to be a good resolution for key estimation [36] and structural analysis [51].

TABLE VII  
F-MEASURE FOR DIFFERENT AVERAGING FACTORS AND CONSTRAINTS. DTW APPROACH EVALUATION WITH DB75

Alg. name	Constr. type	5	10	15	20	25	30	40
Sakoe-Chiba	Global	0.259	0.282	0.327	0.332	0.342	<b>0.355</b>	0.331
Itakura	Global	0.256	0.286	0.362	0.353	0.360	<b>0.395</b>	0.388
Simple DTW	No constr.	0.537	0.606	0.611	0.632	<b>0.638</b>	0.634	0.598
MyersT1	Local	0.647	<b>0.651</b>	0.641	0.643	0.624	0.625	0.577
MyersT2	Local	<b>0.651</b>	0.646	0.617	0.614	0.599	0.566	0.542

2) *Similarity measure between features*: In section III-B we have stated the importance of the similarity employed to compare chroma vectors. Furthermore, we have shown that using a similarity measure that is well correlated with cognitive foundations of musical pitch [46] improves substantially the final system accuracy. When using tonality descriptors, some papers do not specify how a local distance between these feature vectors is computed. They are supposed to assess chroma features' similarity as the rest of studies: with an euclidean-based distance. Since tonality features such as chroma vectors are proven not to be in an euclidean space [52], [53], [54], [55], this assumption seems to be wrong. Furthermore, any method (e.g., a classifier) using distances and concepts just valid for an euclidean space will have the same problem. This is an important issue that will be dealt in the proposed method (section IV).

3) *Chroma transposition*: To account for main key differences, one song is transposed to the tonality of the other one by means of computing a global HPCP for each song (section III-C) and circularly shifting by the OTI (equation 1). This technique has been proven to be more accurate than transposing the song to a reference key by means of a key estimation algorithm. In this case, the use of a less-than-perfect key extraction algorithm degrades the overall identification accuracy. Through the testing of two transposition variants we have pointed out the relevance this fact has in a cover song identification system or in a tonal alignment algorithm.

4) *The use of beat tracking*: We have seen that the DTW approach summarized in section II-B2 could lead to better results without beat tracking information (tables V and VII). Better results for DTW without beat tracking information were also found when comparing against the cross-correlation approach (which uses beat information). We can see this in table IX and in figure 8 (we also provide an extra comparative figure in a separate web page<sup>8</sup>). This is another fact that makes us disregard the use of 'intermediate' processes such as key estimation algorithms and beat tracking systems (citing the two that have been tested here), or chord and melody extraction engines. We feel that this can be a double-edged sword. Due to the fact that all these methods do not have a fully reliable performance<sup>9</sup>, they may decrease the accuracy of a system comprising (at least) one of them. The same argument can be applied to any audio segmentation, chorus extraction, or summarization technique. We can also take a look at state-of-the-art approaches. For instance, common accuracy values

for a chord recognition engine range from 75.5% [56] to 93.3% [57] depending on the method and the considered music material. Also, in this last case, once the chords are obtained, the approach to measure distances between them is still an unsolved issue, involving both some cognitive and musicological concepts that are not fully understood yet. So, errors in these 'intermediate' processes might be added (in case we are using more than one of them), and be propagated to the overall system's identification accuracy (the so called *weakest link* problem).

5) *Alignment procedure*: Several tests have been presented with chroma features DTW alignment. DTW allows us to restrict the alignment (or 'warping') paths to our requirements (section III-E). Consequently, we have tested four 'standard' constraints on these paths (two local and two global constraints). With global constraints we are not considering paths (or alignments) that might be far from the DTW matrix main diagonal. A problem arises when this path can represent a 'correct' alignment (as the example illustrated in figure 4). We have also seen that the accuracy decreases substantially with these constraints. As mentioned in section I, covers can substantially alter the song structure. When this happens, the 'correct' alignment between two covers of the same *canonical song* may be outside of the main DTW matrix diagonal. Therefore, the use of global constraints dramatically decreases the system detection accuracy. These two facts reveal the incorrectness of using a global alignment technique for cover song identification. Regarding local constraints, we have seen that these can help us by reducing 'pathological' warpings that arise when using a small *averaging factor* (table VII). Consequently, this allows us to use much detail in our analysis, and, therefore, to get a better accuracy.

Many systems for cover song identification use a global alignment technique such as DTW or entire song cross-correlation for determining similarity (except the ones that use a summarization, chorus extraction or segmentation technique, which would suffer from the problem of the 'weakest link', cited above). In our opinion, a system considering similarity between song subsequences, and thus, using a local similarity or alignment method, is the only way to cope with strong song structural changes.

#### IV. PROPOSED METHOD

In this section we present a novel method for cover song identification which tries to avoid all the weak points that conventional methods may have and which have been analyzed in previous section. The proposed method uses high-resolution HPCPs (36-bin) as these have been shown to lead to better accuracy (section III-A). To account for key transpositions, the

<sup>8</sup><http://mtg.upf.edu/~jserra/chromabinsimappendix.html>

<sup>9</sup>To account for accuracies of those systems you can visit, e.g., MIREX 2006 wiki page: [http://www.music-ir.org/mirex/2006/index.php/Main\\_Page](http://www.music-ir.org/mirex/2006/index.php/Main_Page) (Accessed 29 Jan. 2008)

OTI transposition method explained in section III-C is used instead of a conventional key finding algorithm. We avoid using any kind of ‘intermediate’ technique as key estimation, chord extraction or beat tracking, as these might degrade the final system identification accuracy (as discussed in section III-F). The method does not employ global constraints, and takes advantage of the improvement given by the local constraints explained in section III-E. Furthermore, it presents relevant differences in two important aspects that boost its accuracy in a dramatic way: it uses a new binary similarity function between chroma features (we have verified the relevance of distance measures in section III-B), and employs a novel local alignment method accounting for structural changes (considering similarity between subsequences, as discussed in section III-F).

A quite resemblant method to the one proposed here is [12]. In there, a chroma-based feature named Polyphonic Binary Feature Vector (PBFV) is adopted, which uses spectral peaks extraction and harmonics elimination. Then, the remaining spectral peaks are averaged across beats and collapsed to a 12-element binary feature vector. This results in a string vector for each analyzed song. Finally, a fast local string search method and a Dynamic Programming (DP) matching are evaluated. The method proposed here also extracts a chroma feature vector using only spectral peaks (HPCP, see section II-A), but we do not do beat averaging, which we find has a detrimental effect in the accuracy of DP algorithms such as Dynamic Time Warping (DTW) (section III-D). Another important difference to the proposed method is the similarity between vectors. In [12], this is computed between binarized vectors, while in the proposed method, what is binarized is the similarity measure, not the vectors themselves (equation 3). Finally, we also think that using an exhaustive alignment method like the one proposed in next section IV-A is also determinant for our final system identification accuracy.

#### A. System description

Figure 5 shows a general block diagram of the system. It comprises four main sequential modules: pre-processing, similarity matrix creation, dynamic programming local alignment (DPLA) and post-processing.

From each pair of compared songs A and B (inputs), we obtain a distance between them (output). Pre-processing comprises HPCP sequence extraction and a global HPCP averaging for each song. Then, one song is transposed to the key of the other one by means of an *Optimal Transposition Index* (OTI). From these two sequences, a binary similarity matrix is then computed. This last is the only input needed for a *Dynamic Programming Local Alignment* (DPLA) algorithm, which calculates a score matrix that gives highest ratings to best aligned subsequences. Finally, in the post-processing step, we obtain a normalized distance between the two processed songs. We now explain these steps in detail.

1) *Pre-processing*: For each song, we extract a sequence of 36-bin HPCP feature vectors as made before, using the same parameters specified in section II-A. An averaging factor of 10 was used as it was found to work well in sections III-D

and III-E. As we are using local constraints for the proposed method, it is not surprising to find a quite similar identification accuracy curve for different values of the averaging factor when comparing the proposed method with the locally constrained DTW algorithms explained in section III-E. In an electronic appendix to this article<sup>10</sup>, the interested reader can find a figure showing the accuracy curves for the proposed method and for DTW with local constraints [45].

A global HPCP vector is computed by averaging all HPCPs in a sequence, and normalizing by its maximum value. With the global HPCPs of two songs ( $\vec{h}_A$  and  $\vec{h}_B$ ), we compute the OTI index, which represents the number of bins that an HPCP needs to circularly shift to have maximal resemblance to the other (see equation 1 in section III-C).

The last operation of the pre-processing block consists in transposing both musical pieces to a common key. This is simply done by circularly shifting each HPCP in the whole sequence of just one song by  $OTI(\vec{h}_A, \vec{h}_B)$  bins (remember we denote musical transposition by superscript  $^{Tr}$ ).

2) *Similarity matrix*: The next step is computing a similarity matrix  $S$  between the obtained pair of HPCP sequences. Notice that the sequences can have different lengths  $n$  and  $m$ , and that, therefore,  $S$  will be an  $n \times m$  matrix. Element  $(i, j)$  of the similarity matrix  $S$ , has the functionality of a local sameness measure between HPCP vectors  $\vec{h}_{A,i}^{Tr}$  and  $\vec{h}_{B,j}$  ( $S_{i,j} = s(\vec{h}_{A,i}^{Tr}, \vec{h}_{B,j})$ ). In our case, this is binary (i.e., only two values are allowed).

We outline some reasons for using a binary similarity measure between chroma features. First, as these features might not be in an euclidean space [46], we would prefer to avoid the computation of an euclidean-based (dis)similarity measure (in general, we think that tonal similarity, and therefore chroma feature distance, is a still far to be understood topic, with many of perceptual and cognitive open issues). Second, using only two values to represent similarity, the possible paths through the similarity matrix become more evident, providing us with a clear notion of where the two sequences agree and where they mismatch (see figure 6 for an example). In addition, binary similarity allows us to operate like many string alignment techniques do: just considering if two elements of the string are the same. With this, we have an expanded range of alignment techniques borrowed from string comparison, DNA or protein sequence alignment, symbolic time series similarity, etc. [32]. Finally, we believe that considering the binary similarity of an HPCP vector might be an easier (or at least more affordable) task to assess than obtaining a reliable graded scale of resemblance between two HPCPs correlated with (sometimes subjective) perceptual similarity.

An intuitive idea to consider when deciding if two HPCP vectors refer to the same tonal root is to keep circularly shifting one of them and to calculate a resemblance index for all possible transpositions. Then, if the transposition that leads to maximal similarity corresponds to less than a semitone (accounting for slight tuning differences), the two HPCP vectors are claimed to be the same. This idea can be formulated

<sup>10</sup><http://mtg.upf.edu/~jserra/chromabinsimappendix.html>

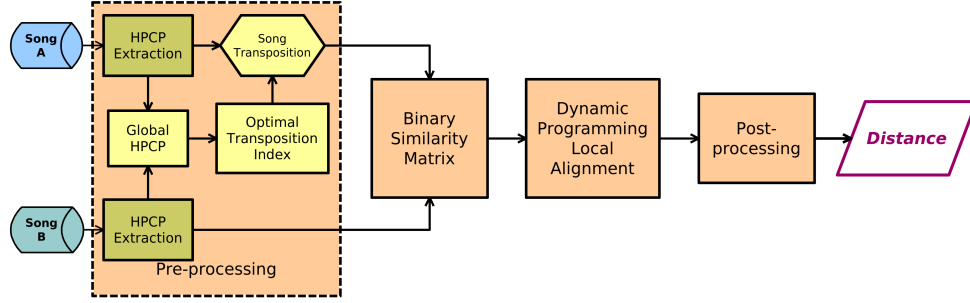


Fig. 5. General block diagram of the system.

in terms of the OTI explained in equation 1. So, as we are using a resolution of a  $1/3$  of a semitone (36 bins), the binary similarity measure between the two vectors is then obtained by:

$$s(\vec{h}_{A,i}^{Tr}, \vec{h}_{B,j}) = \begin{cases} \mu_+ & \text{if } OTI(\vec{h}_{A,i}^{Tr}, \vec{h}_{B,j}) \in \{0, 1, N_H - 1\}, \\ \mu_- & \text{otherwise.} \end{cases} \quad (3)$$

where  $\mu_+$  and  $\mu_-$  are two constants that indicate match or mismatch. These are usually set to a positive and a negative value (e.g., +1 and -1). Empirically, we found that a good choice for  $\mu_+$  and  $\mu_-$  were +1 and -0.9 respectively. Ranges of  $\mu_+$  and  $\mu_-$  between  $\pm 0.7$  and  $\pm 1.25$  resulted in changes smaller than an 5% of the evaluation measures tested. We show two examples of this type of similarity matrix in figure 6.



Fig. 6. Euclidean-based similarity matrix for two covers of the same song (left), OTI-based binary similarity matrix for the same covers (center) and OTI-based binary similarity matrix for two songs that do not share a common tonal progression (right). We can see diagonal white lines in the second plot, while this pattern does not exist in the third. Coordinate units in the horizontal and vertical axes correspond to 1 sec frames.

3) *Dynamic programming local alignment (DPLA)*: A binary similarity matrix  $S$  is the only input to our DPLA algorithm. In section III-E we have seen that using global constraints and, thus, forcing warping paths to be around the alignment matrix main diagonal, had a detrimental effect in final system accuracy. Instead, the use of local constraints [50] can help us preventing ‘pathological warpings’ and just admitting certain ‘logical’ tempo changes. Also, in section III-F, it has been discussed the suitability of performing a local alignment to overcome strong song structure changes (i.e., to check all possible subsequences). The Smith-Waterman algorithm [58] is a well-known algorithm for performing local sequence alignment in Molecular Biology. It was originally

designed for determining similar regions between two nucleotide or protein sequences. Instead of looking at the total sequence, the Smith-Waterman algorithm compares segments of all possible lengths and optimizes the similarity measure.

So, in the same manner as the Smith-Waterman algorithm does, we create an  $(n + 1) \times (m + 1)$  alignment matrix  $H$  through a recursive formula, that, in addition, incorporates some local constraints:

$$H_{i,j} = \max \begin{cases} H_{i-1,j-1} + S_{i-1,j-1} - \delta(S_{i-2,j-2}, S_{i-1,j-1}) \\ H_{i-2,j-1} + S_{i-1,j-1} - \delta(S_{i-3,j-2}, S_{i-1,j-1}) \\ H_{i-1,j-2} + S_{i-1,j-1} - \delta(S_{i-2,j-3}, S_{i-1,j-1}) \\ 0 \end{cases} \quad (4)$$

for  $4 \leq i \leq n+1$  and  $4 \leq j \leq m+1$ . Each  $S_{i,j}$  corresponds to the value of the binary similarity matrix  $S$  at element  $(i, j)$ , and  $\delta()$  denotes a penalty for a gap opening or extension. This latter value is set to 0 if  $S_{i-1,j-1} > 0$  (no gap between  $S_{i-1,j-1}$  and either  $S_{i-2,j-2}$ ,  $S_{i-3,j-2}$  or  $S_{i-2,j-3}$ ), or to a positive value if  $S_{i-1,j-1} \leq 0$ . More concretely:

$$\delta(a, b) = \begin{cases} 0 & \text{if } b > 0 \text{ (no gap)} \\ c_1 & \text{if } b \leq 0 \text{ and } a > 0 \text{ (gap opening)} \\ c_2 & \text{if } b \leq 0 \text{ and } a \leq b \text{ (gap extension)} \end{cases} \quad (5)$$

Good values were empirically found to be  $c_1 = 0.5$  for a gap opening, and  $c_2 = 0.7$  for a gap extension. Small variability of the evaluation measures was shown for  $c_1, c_2$  values between 0.3 and 1. We used the songs in DB90 for empirically estimating these parameters and then evaluated the method with DB2053 (see section IV-B).

Values of  $H$  can be interpreted considering that  $H_{i,j}$  is the maximum similarity of two segments ending in  $h_{A,i-1}^{Tr}$  and  $h_{B,j-1}$  respectively. The zero is included to prevent negative similarity, indicating no similarity up to  $h_{A,i-1}^{Tr}$  and  $h_{B,j-1}$ . The first 3 rows and columns of  $H$  can be initialized to have a 0 value.

An example of the resultant matrix  $H$  is shown in figure 7. We clearly observe two local alignment traces, which correspond to two highly resemblant sections between two



versions of the same song (from  $H_{150,25}$  to  $H_{250,100}$  and from  $H_{280,25}$  to  $H_{400,100}$ , where sub-indices respectively denote rows and columns).

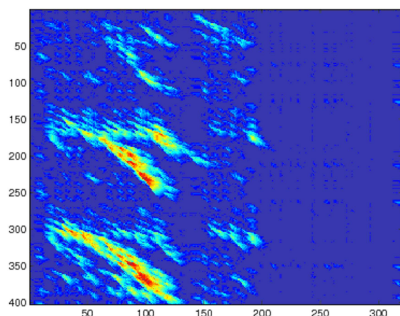


Fig. 7. Example of a local alignment matrix  $H$  between two covers. It can be seen that the two songs do not entirely coincide (just in two fragments), and that, mainly, their respective second halves are completely different. Coordinate units in the horizontal and vertical axes correspond to 1 sec averaging across frames.

4) *Post-processing*: In the last step of the method, only the best local alignment in  $H$  is considered. This means that the score determining the local subsequence similarity between two HPCP sequences, and, therefore, what we consider to be the similarity between two songs, corresponds to the value of  $H$ 's highest peak:

$$\text{Score}(HPCP_A^{Tr}, HPCP_B) = \max\{H_{i,j}\} \quad (6)$$

for any  $i, j$  such that  $1 \leq i \leq n + 1$  and  $1 \leq j \leq m + 1$ .

Finally, to obtain a dissimilarity value that is independent of song duration, the score is normalized by the compared song lengths [45] and the inverse is taken:

$$d(\text{song}_A, \text{song}_B) = \frac{n + m}{\text{Score}(HPCP_A^{Tr}, HPCP_B)} \quad (7)$$

where  $n$  and  $m$  are the respective lengths for songs A and B.

## B. Evaluation

We now display the results corresponding to the evaluation of our method. This has been made with the music collection presented in section II-C and within the framework of the MIREX 2008 Audio Cover Song Identification contest as well. As the databases used in this part of the paper may have more than 5 covers per set, the first 10 retrieved items were considered for evaluation.

Firstly, as we have proposed a new distance measure between chroma features, we provide results for a comparison between common distance measures and the proposed OTI-based binary distance in table VIII. To perform this comparison, we have thresholded common distance measures and applied the same DPLA algorithm (with the same parameters) to all of them. Several thresholds were tested for each distance

in order to determine the ones leading to best identification accuracy. We observe that OTI-based binary similarity matrix outperforms other binary similarity matrices obtained through thresholding common similarity measures between chroma features. In the case of these last measures, best identification accuracy values for different thresholds tested are shown.

TABLE VIII  
IDENTIFICATION ACCURACY FOR DPLA ALGORITHM WITH 5 DIFFERENT BINARY SIMILARITY MATRICES AS INPUT. EVALUATION DONE WITH DB2053

Distance used	F-measure	R <sub>10</sub>
Dot product	0.132	0.136
Euclidean distance	0.218	0.216
Cosine similarity	0.221	0.219
Correlation	0.239	0.247
OTI-based similarity	0.601	0.576

We next show the general evaluation results corresponding to our personal music collection. Within these, we compare identification accuracy between the proposed method and the best variants of the cross-correlation and DTW methods tested. In table IX we report the F-measure values for the three different databases presented. Recall is shown in figure 8. In there, we plot an average Recall figure for all the implemented systems (best variants). Vertical axis represents Recall and horizontal axis represents different percentages of the retrieved answer. As this was set to a maximum length of 10, the numbers represent 0 answers (giving a Recall of 0), 1 answer, 2 answers and so forth. We can see that with the newly proposed method the accuracy is around 58% of correctly retrieved songs within the first 10 retrieved answers. This value is highly superior to the accuracies achieved for the best versions of the cross-correlation and DTW methods that we could implement (around 20 and 40 percent respectively), and is very far from the the baseline corresponding to just guessing by chance, which is lower than 0.3%.

TABLE IX  
F-MEASURE FOR THE PROPOSED METHOD, THE DTW AND THE CROSS-CORRELATION APPROACHES. PARAMETERS FOR THE CROSS-CORRELATION AND THE DTW METHODS WERE ADJUSTED ACCORDING TO THE BEST VALUES AND VARIANTS FOUND IN SECTION III

Method	DB75	DB330	DB2053
Cross-correlation	0.638	0.348	0.169
DTW	0.651	0.485	0.399
Proposed method	0.868	0.688	0.601

If we take a look to MIREX 2007 contest data (where we participated with this algorithm), we observe that our system was the best performing one with a substantial difference to others [59]. A total of 8 different algorithms were presented to the MIREX 2007 Audio Cover Song task. Table X shows the overall summary results obtained<sup>11</sup>. The present algorithm (SG, first column) performed the best in all considered evaluation measures, reaching an average accuracy of 5.009 of correctly identified covers within the 10 first retrieved elements ( $MNCI_{10}$ ) and a Mean Average Precision ( $MAP$ ) of 0.521.

<sup>11</sup>See the complete results and details about the evaluation procedure at [http://www.music-ir.org/mirex/2007/index.php/Audio\\_Cover\\_Song\\_Identification\\_Results](http://www.music-ir.org/mirex/2007/index.php/Audio_Cover_Song_Identification_Results) (Accessed 29 Jan. 2008)

TABLE X

RESULTS FOR MIREX 2007 AUDIO COVER SONG TASK. ACCURACY MEASURES EMPLOYED WERE THE TOTAL NUMBER OF COVERS IDENTIFIED WITHIN THE FIRST 10 ANSWERS ( $TNCI_{10}$ ), THE MEAN NUMBER OF COVERS IDENTIFIED WITHIN THE 10 FIRST ANSWERS ( $MNCI_{10}$ ), THE MEAN OF AVERAGE PRECISION ( $MAP$ ) AND THE AVERAGE RANK OF THE FIRST CORRECTLY IDENTIFIED COVER ( $RANK_1$ ). CLOCK TIME MEASURES ARE REPORTED ON THE LAST LINE OF THE TABLE (NUMBER OF USED THREADS IN BRACKETS). VALUES FOR THE ALGORITHM PRESENTED HERE ARE SHOWN IN THE FIRST COLUMN (SG)

Measure	Range	SG	EC	JB	JEC	KL1	KL2	KP	IM
$TNCI_{10}$	[0-3300]	<b>1653</b>	1207	869	762	425	291	190	34
$MNCI_{10}$	[0-10]	<b>5.009</b>	3.658	2.633	2.309	1.288	0.882	0.576	0.103
$MAP$	[0-1]	<b>0.521</b>	0.330	0.267	0.238	0.13	0.086	0.061	0.017
$Rank_1$	[0-1000]	<b>9.367</b>	13.994	29.527	22.209	57.542	51.094	46.539	97.470
Runtime	[HH:MM]	<b>01:37(1)</b>	04:28(5)	04:32(8)	00:47(8)	10:45(8)	02:37(1)	03:51(1)	02:04(1)

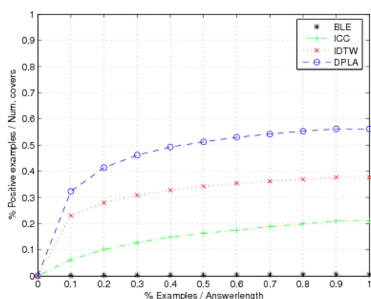


Fig. 8. Average Recall figures comparing the proposed approach (blue circles) with the cross-correlation (green sum signs) and the DTW (red crosses) methods for DB2053. Parameters for the cross-correlation and the DTW methods compared were adjusted according to the best values found in section III. A base-line identification accuracy (BLE) is also plotted (black bottom asterisks).

Furthermore, the next best performing system reached an  $MNCI_{10}$  of 3.658 and a  $MAP$  of 0.330, which represents a substantial difference to the one proposed in this paper (57.88% superior in terms of  $MAP$ ). In addition, statistical significance tests showed that the results for the system were significantly better than those of the other six systems presented in the contest.

A basic error analysis [45] shows that the best identified covers are “A forest”, originally performed by The Cure and “Let it be”, originally performed by The Beatles. Other correctly classified items are “Yesterday”, “Dont let me down” and “We can work it out”, all originally performed by The Beatles and “How insensitive” (Vinicius de Moraes). This high amount of Beatles’ songs within the better classified items can be due to the fact that there were many Beatles’ cover sets (e.g., 14 out of 30 in DB330), but it can also be justified considering the clear simplicity and definition of their tonal progressions, that, in comparison with other more elaborated pieces (e.g., “Over the rainbow” performed by Judy Garland), leads to better identification. Within this set of better identified covers there are several examples of structural changes and tempo deviations. In the electronic appendix<sup>12</sup>, we provide a confusion matrix with labels corresponding to cover sets (rows and columns).

We detected that there were some songs, such as “Eleanor

<sup>12</sup><http://mtg.upf.edu/~jserra/chromabinsimappendix.html>

Rigby” and “Get Back”, that caused ‘confusion’ more or less with all the queries made. One explanation for this might be that these two songs are built over a very simple chord progression involving just two chords: the tonic and the mediant (e.g., C and Em for a C major key) for the former, and the tonic and the subdominant (e.g., C and F for a C major key) for the latter. So, as they rely half of the time in the tonic chord, any song being compared to them will share half of the tonal progression. Other poorly classified items are “The battle of Epping forest” (Genesis) or “Stairway to heaven” (Led Zeppelin). Checking their wrongly associated covers, we find that, most of the time, the alignment, the similarity measure and the transposition are performing correctly according to the features extracted. Thus, we have the intuition that the tonal progression might not be enough for some kinds of covers. This does not mean that HPCPs could be sensitive to timbre or other facets of the musical pieces. On the contrary, we are able to detect many covers that have a radical change in the instrumentation, which we think it is due to the capacity of HPCPs to filter timbre out.

An interesting misclassification appears with “No woman no cry”, originally performed by Bob Marley. These covers are associated more than 1/3 of the times with the song “Let it be” (The Beatles). When we analyzed the harmonic progression of both songs, we discovered that they share the same chords in different parts of the theme (C - G - Am - F). Thus, this might be a logical misclassification using chroma features. Another source of frequent confusion is the classical harmonic progression I - IV - I or I - V - IV - I, which many songs share.

## V. CONCLUSIONS

In this paper we have devised a new method for audio signal comparison focused on cover song identification that by large outperforms state-of-the-art system. This has been achieved after experimenting with many proposed techniques and variants, and testing their effect in final identification accuracy, which also was one of the main objectives in writing this article.

We have first presented our test framework and the two state-of-the-art methods that we have used in further experiments. The performed analysis has focused on several variants that could be taken for these two methods (and, in general, for any method based on chroma descriptors): (a) the chroma features resolution - section III-A; (b) the local cost function (dissimilarity measure) between chroma features - section

III-B; (c) the effect of using key transposition methods - section III-C; and (d) the use of a beat tracking algorithm to obtain a tempo-independent representation of the chroma sequence - section III-D. In addition, as DTW is a well known and extensively used technique, we tested two underexplored variants of it, apart from the simple one mentioned in section II-B2: DTW with global and with local constraints (section III-E). The results of these cross-validated experiments have been summarized in section III-F.

Finally, we have presented a new cover song identification system that takes advantage of the results found and that has been proven, using different evaluation measures and contexts, to work significantly better than other state-of-the-art methods. Although cover song identification is still a relatively new research topic, and systems dealing with this task can be further improved, we think that the work done and the method presented here represent an important milestone.

#### ACKNOWLEDGEMENTS

The authors would like to thank their colleagues and staff at the Music Technology Group (UPF) for their support and encouragement, especially Graham Coleman for his review and proofreading. Furthermore, the authors wish to thank the anonymous reviewers for very helpful comments.

#### REFERENCES

- [1] R. Witmer and A. Marks, "Cover", *Grove Music Online*, L. Macy, Ed. Oxford University Press, 2006, (Accessed 25 Oct. 2007), <http://www.grovemusic.com>.
- [2] S. Strunk, "Harmony", *Grove Music Online*, L. Macy, Ed. Oxford University Press, 2006, (Accessed 26 Nov. 2007), <http://www.grovemusic.com>.
- [3] S. Dalla Bella, I. Peretz, and N. Aronoff, "Time course of melody recognition: A gating paradigm study," *Perception and Psychophysics*, vol. 7, no. 65, pp. 1019–1028, 2003.
- [4] M. D. Schulkind, R. J. Posner, and D. C. Rubin, "Musical features that facilitate melody identification: How do you know it's your song when they finally play it?" *Music Perception*, vol. 21, no. 2, pp. 217–249, 2003.
- [5] N. Hu, R. B. Dannenberg, and G. Tzanetakis, "Polyphonic audio matching and alignment for music retrieval," *IEEE Workshop on Apps. of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 185–188, 2003.
- [6] N. H. Adams, N. A. Bartsch, J. B. Shifrin, and G. H. Wakefield, "Time series alignment for music information retrieval," *Int. Symp. on Music Information Retrieval (ISMIR)*, pp. 303–310, 2004.
- [7] M. Casey and M. Slaney, "The importance of sequences in musical similarity," *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 5, pp. V–V, May 2006.
- [8] W. H. Tsai, H. M. Yu, and H. M. Wang, "A query-by-example technique for retrieving cover versions of popular songs with similar melodies," *Int. Symp. on Music Information Retrieval (ISMIR)*, pp. 183–190, 2005.
- [9] M. Marolt, "A mid-level melody-based representation for calculating audio similarity," *Int. Symp. on Music Information Retrieval (ISMIR)*, pp. 280–285, 2006.
- [10] Ö. Izmirli, "Tonal similarity from audio using a template based attractor model," *Int. Symp. on Music Information Retrieval (ISMIR)*, pp. 540–545, 2005.
- [11] J. P. Bello, "Audio-based cover song retrieval using approximate chord sequences: testing shifts, gaps, swaps and beats," *Int. Symp. on Music Information Retrieval (ISMIR)*, pp. 239–244, September 2007.
- [12] H. Nagano, K. Kashino, and H. Murase, "Fast music retrieval using polyphonic binary feature vectors," *IEEE Int. Conf. on Multimedia and Expo (ICME)*, vol. 1, pp. 101–104, 2002.
- [13] M. Müller, F. Kurth, and M. Clausen, "Audio matching via chroma-based statistical features," *Int. Symp. on Music Information Retrieval (ISMIR)*, pp. 288–295, 2005.
- [14] M. Casey and M. Slaney, "Song intersection by approximate nearest neighbor search," *Int. Symp. on Music Information Retrieval (ISMIR)*, pp. 144–149, October 2006.
- [15] E. Gómez, B. S. Ong, and P. Herrera, "Automatic tonal analysis from music summaries for version identification," *Conv. of the Audio Engineering Society (AES)*, October 2006.
- [16] D. P. W. Ellis and G. E. Poliner, "Identifying cover songs with chroma features and dynamic programming beat tracking," *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, pp. 1429–1432, April 2007.
- [17] A. Klapuri, "Signal processing methods for the automatic transcription of music," Ph.D. dissertation, Tampere University of Technology, Finland, April 2004.
- [18] M. Goto, "A real-time music-scene-description system: predominant-f0 estimation for detecting melody and bass lines in real-world audio signals," *Speech Communication*, vol. 43, no. 4, pp. 311–329, September 2004.
- [19] G. E. Poliner, D. P. W. Ellis, A. Ehmann, E. Gómez, S. Streich, and B. S. Ong, "Melody transcription from music audio: approaches and evaluation," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, pp. 1247–1256, 2007.
- [20] A. Sheh and D. P. W. Ellis, "Chord segmentation and recognition using em-trained hidden markov models," *Int. Symp. on Music Information Retrieval (ISMIR)*, pp. 183–189, 2003.
- [21] C. A. Harte and M. B. Sandler, "Automatic chord identification using a quantized chromagram," *Conv. of the Audio Engineering Society (AES)*, pp. 28–31, 2005.
- [22] T. Fujishima, "Realtime chord recognition of musical sound: a system using common lisp music," *Int. Computer Music Conference (ICMC)*, pp. 464–467, 1999.
- [23] G. Tzanetakis, "Pitch histograms in audio and symbolic music information retrieval," *Int. Symp. on Music Information Retrieval (ISMIR)*, pp. 31–38, 2002.
- [24] S. Paws, "Musical key extraction from audio," *Int. Symp. on Music Information Retrieval (ISMIR)*, pp. 96–99, 2004.
- [25] E. Gómez, "Tonal description of music audio signals," Ph.D. dissertation, MTG, Universitat Pompeu Fabra, Barcelona, Spain, 2006, <http://mtg.upf.edu/~egomez/thesis>.
- [26] R. B. Dannenberg and N. Hu, "Pattern discovery techniques for music audio," *Int. Symp. on Music Information Retrieval (ISMIR)*, pp. 63–70, 2002.
- [27] N. A. Bartsch and G. H. Wakefield, "To catch a chorus: using chroma-based representations for audio thumbnailing," *IEEE Workshop on Apps. of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 15–18, 2001.
- [28] M. Goto, "A chorus-section detection method for musical audio signals and its application to a music listening station," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1783–1794, September 2006.
- [29] M. Müller, *Information Retrieval for Music and Motion*. Springer, 2007.
- [30] L. R. Rabiner and B. H. Juang, *Fundamentals of speech recognition*. Englewood Cliffs, NJ: Prentice, 1993.
- [31] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet Physics Doklady*, vol. 10, pp. 707–710, 1966.
- [32] D. Gusfield, *Algorithms on strings, trees and sequences: computer sciences and computational biology*. Cambridge University Press, 1997.
- [33] P. Cano, M. Kaltenbrunner, O. Mayor, and E. Batlle, "Statistical significance in song-spotting in audio," *Int. Symp. on Music Information Retrieval (ISMIR)*, pp. 77–79, 2001.
- [34] R. L. Kline and E. P. Glinert, "Approximate matching algorithms for music information retrieval using vocal input," *ACM Multimedia*, pp. 130–139, 2003.
- [35] A. Gionis, P. Indyk, and R. Motwani, "Similarity search in high dimensions via hashing," *Very Large Databases Journal*, pp. 518–529, 1999.
- [36] E. Gómez and P. Herrera, "Estimating the tonality of polyphonic audio files: cognitive versus machine learning modelling strategies," *Int. Symp. on Music Information Retrieval (ISMIR)*, pp. 92–95, 2004.
- [37] —, "The song remains the same: identifying versions of the same song using tonal descriptors," *Int. Symp. on Music Information Retrieval (ISMIR)*, pp. 180–185, 2006.
- [38] B. S. Ong, E. Gómez, and S. Streich, "Automatic extraction of musical structure using pitch class distribution features," *Workshop on Learning the Semantics of Audio Signals (LSAS)*, pp. 53–65, 2006.

- [39] H. Purwins, "Proles of pitch classes. circularity of relative pitch and key: experiments, models, computational music analysis, and perspectives." Ph.D. dissertation, Berlin University of Technology, Germany, 2005.
- [40] D. Huron, "Scores from the ohio state university cognitive and systematic musicology laboratory - bach well-tempered clavier fugues, book ii," Online: <http://kern.ccarh.org/cgi-bin/ksbrowse?l=osu/classical/bach/wtc-2>, 1994, (Last access Jan. 2008).
- [41] M. E. P. Davies and P. Brossier, "Beat tracking towards automatic musical accompaniment," *Conv. of the Audio Engineering Society (AES)*, May 2005.
- [42] P. Brossier, "Automatic annotation of musical audio for interactive applications," Ph.D. dissertation, Queen Mary, London, 2007.
- [43] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. ACM Press Books, 1999.
- [44] J. Serrà, "A qualitative assessment of measures for the evaluation of a cover song identification system," *Int. Symp. on Music Information Retrieval (ISMIR)*, pp. 319–322, September 2007.
- [45] —, "Music similarity based on sequences of descriptors: tonal features applied to audio cover song identification," Master's thesis, MTG, Universitat Pompeu Fabra, Barcelona, Spain, 2007.
- [46] C. L. Krumhansl, *Cognitive foundations of musical pitch*. New York: Oxford University Press, 1990.
- [47] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequences of two proteins," *Journal of Molecular Biology*, vol. 48, pp. 443–453, 1970.
- [48] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimisation for spoken word recognition," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 26, pp. 43–49, 1978.
- [49] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 23, pp. 52–72, 1975.
- [50] C. Myers, "A comparative study of several dynamic time warping algorithms for speech recognition," Master's thesis, Massachusetts Institute of Technology (MIT), USA, 1980.
- [51] B. S. Ong, "Structural analysis and segmentation of music signals," Ph.D. dissertation, MTG, Universitat Pompeu Fabra, Barcelona, Spain, 2007.
- [52] R. N. Shepard, "Structural representations of musical pitch," *The Psychology of Music*, 1982.
- [53] D. Lewis, *Generalized musical intervals and transformations*. Newhaven: Yale University Press, 1987.
- [54] R. Cohn, "Neo-riemannian operations, parsimonious trichords, and their tonnetz representations," *Journal of Music Theory*, vol. 1, no. 41, pp. 1–66, 1997.
- [55] E. Chew, "Towards a mathematical model of tonality," Ph.D. dissertation, Massachusetts Institute of Technology (MIT), USA, 2000.
- [56] J. P. Bello and J. Pickens, "A robust mid-level representation for harmonic content in music signals," *Int. Symp. on Music Information Retrieval (ISMIR)*, pp. 304–311, 2005.
- [57] K. Lee and M. Slaney, "Automatic chord recognition using an hmm with supervised learning," *Int. Symp. on Music Information Retrieval (ISMIR)*, pp. 133–137, 2006.
- [58] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *Journal of Molecular Biology*, vol. 147, pp. 195–197, 1981.
- [59] J. Serrà and E. Gómez, "A cover song identification system based on sequences of tonal descriptors," *MIREX extended abstract*, 2007.



**Joan Serrà** obtained both the degrees of Telecommunications and Electronics (sound and image specialization) at Enginyeria la Salle, Universitat Ramon Llull (URL), Barcelona, Spain, in 2002 and 2004 respectively. Further studies were focused in digital audio signal processing and audio recording and engineering (Audiovisual Technologies department, URL). After that, he worked in the R+D department of Music Intelligence Solutions Inc., developing patented approaches to music and visual media discovery. In 2006-07, he did the MSc in

Information, Communication and Audiovisual Media Technologies (TICMA) at Universitat Pompeu Fabra (UPF), Barcelona, Spain. He is currently a researcher and a PhD student at the Music Technology Group (MTG) of the UPF. He has also been a semi-professional musician for more than 10 years. His research interests include (but are not limited to): machine learning, music perception and cognition, time series analysis, signal processing, data visualization, dimensionality reduction and information retrieval.



**Emilia Gómez** is a post-doctoral researcher at the Music Technology Group (MTG) of the Universitat Pompeu Fabra (UPF). She graduated as a Telecommunication Engineer specialized in Signal Processing at the Universidad de Sevilla. Then, she received a DEA in Acoustics, Signal Processing and Computer Science applied to Music (ATIAM) at the IRCAM, Paris. In July 2006, she completed her PhD in Computer Science and Digital Communication at the UPF, on the topic of Tonal Description of Music Audio Signals. During her doctoral studies, she was a visiting researcher at the Signal and Image Processing (TSI) group of the École Nationale Supérieure de Télécommunications (ENST), Paris and at the Music Acoustics Group (TMH) of the Stockholm Institute of Technology, KTH. She has been involved in several research projects funded by the European Commission and the Spanish Ministry of Science and Technology. She also belongs to the Department of Sonology, Higher Music School of Catalonia (ESMUC), where she teaches music acoustics and sound synthesis and processing. Her main research interests are related to music content processing, focusing on melodic and tonal facets, music information retrieval and computational musicology.



**Perfecto Herrera** received the degree in Psychology from the University of Barcelona, Spain, in 1987. He was with the University of Barcelona as a Software Developer and an Assistant Professor. His further studies have focused on sound engineering, audio postproduction, and computer music. Now he is finishing his PhD on Music Content Processing in the Universitat Pompeu Fabra (UPF), Barcelona. He has been working in the Music Technology Group, (UPF) since its inception in 1996, first as the person responsible for the sound laboratory/studio, then as a Researcher. He worked in the MPEG-7 standardization initiative from 1999 to 2001. Then, he collaborated in the EU-IST-funded CUIDADO project, contributing to the research and development of tools for indexing and retrieving music and sound collections. This work was somehow continued and expanded as Scientific Coordinator for the Semantic Interaction with Music Audio Contents (SIMAC) project, again funded by the EU-IST. He is currently the Head of the Department of Sonology, Higher Music School of Catalonia (ESMUC), where he teaches music technology and psychoacoustics. His main research interests are music content processing, classification, and music perception and cognition.



**Xavier Serra** (Barcelona, 1959) is the head of the Music Technology Group of the Universitat Pompeu Fabra in Barcelona. After a multidisciplinary academic education he obtained a PhD in Computer Music from Stanford University in 1989 with a dissertation on the spectral processing of musical sounds that is considered a key reference in the field. His research interests are on the understanding, modeling and generating music through computational approaches. He tries to find a balance between basic and applied research with methodologies from both scientific/technological and humanistic/artistic disciplines. Dr. Serra is very active in promoting initiatives in the field of Sound and Music Computing at the international level, being editor and reviewer of a number of international journals, conferences and programs of the European Commission, and giving lectures on current and future challenges of the field. He is the principal investigator of more than 10 major research projects funded by public and private institutions, the author of 31 patents and has published more than 40 research articles.

Laurier, C., Meyers, O., Serrà, J., Blech, M., **Herrera, P.**, Serra, X. (2010). "Indexing Music by Mood: Design and Integration of an Automatic Content-based Annotator". *Multimedia Tools and Applications*. 48(1), 161-184.

DOI: <https://doi.org/10.1007/s11042-009-0360-2>

ISSN: 1380-7501

Online ISSN: 1573-7721



## Indexing music by mood: design and integration of an automatic content-based annotator

Cyril Laurier · Owen Meyers · Joan Serrà ·  
Martin Blech · Perfecto Herrera · Xavier Serra

© Springer Science + Business Media, LLC 2009

**Abstract** In the context of content analysis for indexing and retrieval, a method for creating automatic music mood annotation is presented. The method is based on results from psychological studies and framed into a supervised learning approach using musical features automatically extracted from the raw audio signal. We present here some of the most relevant audio features to solve this problem. A ground truth, used for training, is created using both social network information systems (wisdom of crowds) and individual experts (wisdom of the few). At the experimental level, we evaluate our approach on a database of 1,000 songs. Tests of different classification methods, configurations and optimizations have been conducted, showing that Support Vector Machines perform best for the task at hand. Moreover, we evaluate the algorithm robustness against different audio compression schemes. This fact, often neglected, is fundamental to build a system that is usable in real conditions. In addition, the integration of a fast and scalable version of this technique with the European Project PHAROS is discussed. This real world application demonstrates the usability of this tool to annotate large-scale databases. We also report on a user evaluation in the context of the PHAROS search engine, asking people about the utility, interest and innovation of this technology in real world use cases.

**Keywords** Music information retrieval · Mood annotation · Content-based audio · Social networks · User evaluation

### 1 Introduction

Psychological studies have shown that emotions conveyed by music are objective enough to be valid for mathematical modeling [4, 13, 24, 32]. Moreover, Vieillard et al. [43] demonstrated that within the same culture, the emotional responses to music could be highly consistent. All these results indicate that modeling emotion or mood in music is feasible.

---

C. Laurier (✉) · O. Meyers · J. Serrà · M. Blech · P. Herrera · X. Serra  
Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain  
e-mail: cyril.laurier@upf.edu

In the past few years, research in content-based techniques has been trying to solve the problem of tedious and time-consuming human indexing of audiovisual data. In particular, Music Information Retrieval (MIR) has been very active in a wide variety of topics such as automatic transcription or genre classification [5, 29, 41]. Recently, classification of music mood has become a matter of interest, mainly because of the close relationship between music and emotions [1, 19].

In the present paper, we present a robust and efficient mood annotator that automatically estimates the mood of a piece of music, directly from the raw audio signal. We achieve this task by using a supervised learning method. In Section 2, we report on related works in classification of music mood. In Section 3, we detail the method and the results we achieved. In Section 4, we describe the integration of this technique in the PHAROS project (Platform for searchING of Audiovisual Resources across Online Spaces). In Section 5, we present the protocol and results of a user evaluation. Finally, in Section 6, we discuss future works.

## 2 Scientific background

Although there exist several studies dealing with automatic content-based mood classification (such as [4, 26, 37, 47]), almost every work differs in the way that it represents the mood concepts. Similar to psychological studies, there is no real agreement on a common model [16]. Some consider a categorical representation based on mutually exclusive basic emotions such as “happiness”, “sadness”, “anger”, “fear” and “tenderness” [19, 26, 36, 39], while others prefer a multi-labeling approach (i.e., using a rich set of adjectives that are not mutually exclusive) like Wieczorkowska [45]. The latter is more difficult to evaluate since they consider many categories. The basic emotion approach gives simple but relatively satisfying results, around 70–90% of correctly classified instances, depending on the data and the number of categories chosen (usually between 3 and 5). Li and Ogihara [22] extract timbre, pitch and rhythm features from the audio content to train Support Vector Machines (SVMs). They consider 13 categories, 11 from the ones proposed in Farnsworth [10] plus 2 additional ones. However, the results are not that convincing, obtaining low average precision (0.32) and moderate recall (0.54). This might be due to the small dataset labeled by only one person and to the large number of categories they chose. Conversely, it is very advisable to use few categories and a ground truth annotated by hundreds of people (see Section 3.1).

Other works use the dimensional representation (modeling emotions in a space), like Yang [47]. They model the problem with Thayer’s arousal-valence<sup>1</sup> emotion plane [40] and use a regression approach (Support Vector Regression) to learn each of the two dimensions. They extract mainly spectral and tonal descriptors together with loudness features. The overall results are very encouraging and demonstrate that a dimensional approach is also feasible. In another work, Mandel et al. [27] describe an active learning system using timbre features and SVMs, which learns according to the feedback given by the user. Moreover, the algorithm chooses the examples to be labeled in a smart manner, reducing the amount of data needed to build a model, and has an accuracy equivalent to that of state-of-the-art methods.

Comparing evaluations of these different techniques is an arduous task. With the objective to evaluate different algorithms within the same framework, MIREX (Music Information Retrieval Evaluation eXchange) [8] organized a first task on Audio Mood Classification in 2007.<sup>2</sup>

<sup>1</sup> In psychology, the term valence describes the attractiveness or aversiveness of an event, object or situation. For instance happy and joy have a positive valence and anger and fear a negative valence.

<sup>2</sup> [http://www.music-ir.org/mirex2007/index.php/Audio\\_Music\\_Mood\\_Classification](http://www.music-ir.org/mirex2007/index.php/Audio_Music_Mood_Classification)



MIREX is a reference in the MIR community that provides a solid evaluation of current algorithms in different tasks. The MIREX approach is similar to the Text Retrieval Conference (TREC)<sup>3</sup> approach to the evaluation of text retrieval systems, or TREC-VID<sup>4</sup> for video retrieval. For the Audio Mood Classification task, it was decided to model the mood classification problem with a categorical representation in mood clusters (a word set defining the category). Five mutually exclusive mood clusters were chosen (i.e, one musical excerpt could only belong to one mood cluster). In that aspect, it is similar to a basic emotion approach, because the mood clusters are mutually exclusive. They asked human evaluators to judge a collection of 1,250 30-second excerpts (250 in each mood cluster). The resulting human-validated collection consisted of 600 30-second clips in total. The best results approached 60% of accuracy [14, 18]. In Table 1, we show the categories used and the results of different algorithms, including our submitted algorithm [18] (noted CL). One should note that the accuracies from the MIREX participants are lower than those found in most of the existing literature. This is probably due to a semantic overlap between the different clusters [14]. Indeed, if the categories are mutually exclusive, the category labels have to be chosen carefully.

Performing a statistical analysis on this data with the Tukey-Kramer Honestly Significantly Differently method (TK-HSD) [2], the MIREX organizers found that our algorithm had the first rank across all mood clusters despite its average accuracy being the second highest [14]. Another interesting fact from this evaluation is that, looking at all the submissions, the most accurate algorithms were using SVMs. The results of the MIREX task show that our audio feature extraction and classification method are state-of-the-art. Thus, to create a new music mood annotator, even though we tried different classification methods, we focused on the optimization of Support Vector Machines [3]. Moreover, we especially focused on using a relevant taxonomy and on finding an efficient and original method to create a reliable ground truth.

### 3 Method

To classify music by mood, we frame the problem as an audio classification problem using a supervised learning approach. We consider unambiguous categories to allow for a greater understanding and agreement between people (both human annotators and end-users). We build the ground truth to train our system on both social network knowledge (wisdom of crowds) and experts validation (wisdom of the few). Then we extract a rich set of audio features that we describe in Section 3.2. We employ standard feature selection and classification techniques and we evaluate them in Section 3.3. Once the best algorithm is chosen, we evaluate the contribution of each descriptor in 3.5 and the robustness of the system as reported in Section 3.4. In Fig. 1, we show a general block diagram of the method.

#### 3.1 Ground truth from wisdom of crowds and wisdom of the few

For this study we use a categorical approach to represent the mood. We focus on the following categories: *happy*, *sad*, *angry*, and *relaxed*. We decided to use these categories because these moods are related to basic emotions from psychological theories (reviewed in [15]) and they cover the four quadrants of the 2D representation from Russell [34] with valence and arousal dimensions (see Fig. 2).

<sup>3</sup> <http://trec.nist.gov/>

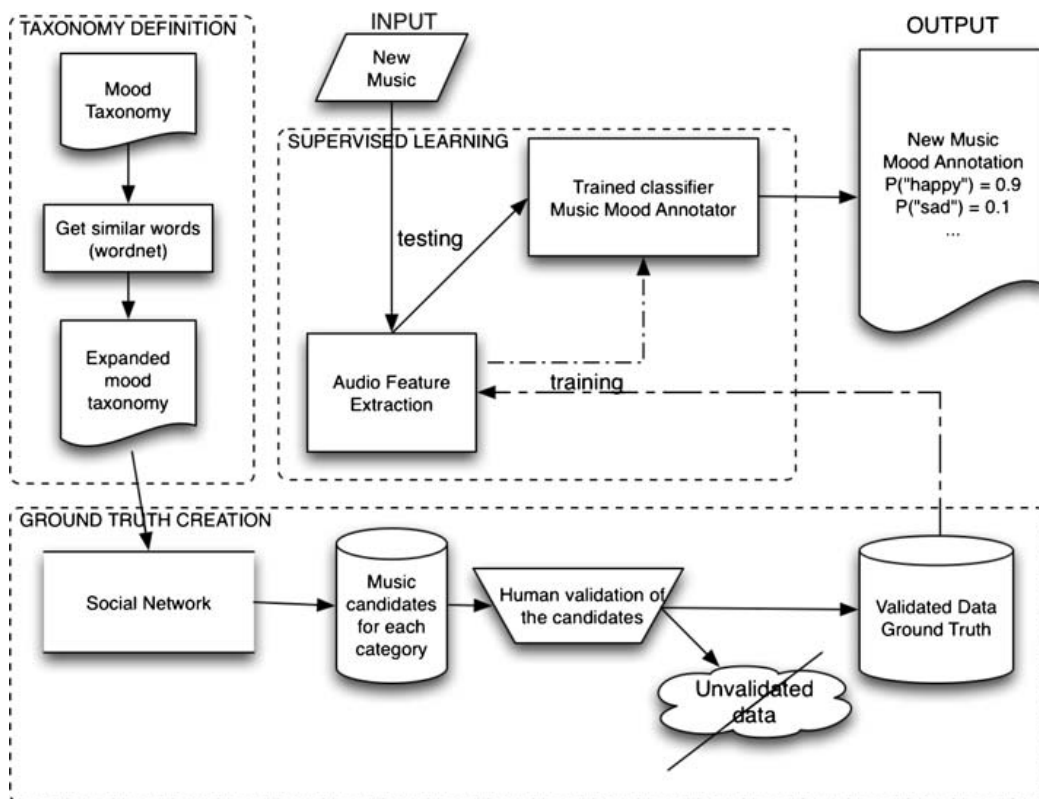
<sup>4</sup> <http://www-nlpir.nist.gov/projects/trecvid/>

**Table 1** Extract from the Audio Mood Classification task results, MIREX 2007. Mean accuracies in percentage over a 3-fold Cross Validation. Comparison of our submitted algorithm (CL [18]), with the other top competitors (GT [42], TL [23], ME [28]). We used several of the audio features presented later in this paper and SVMs

Mood Clusters	CL	GT	TL	ME
rowdy,rousing,confident,boisterous,passionate	45.83%	42.50%	52.50%	51.67%
amiable,good natured,sweet,fun,rollicking,cheerful	50.00%	53.33%	49.17%	45.83%
literate,wistful,bittersweet,autumnal,brooding,poignant	82.50%	80.00%	75.00%	70.00%
witty,humorous,whimsical,wry,campy,quirky,silly	53.33%	51.67%	52.50%	55.00%
volatile,fiery,visceral,aggressive,tense/anxious,intense	70.83%	80.00%	69.17%	66.67%
Mean accuracy	60.50%	61.50%	59.67%	57.83%

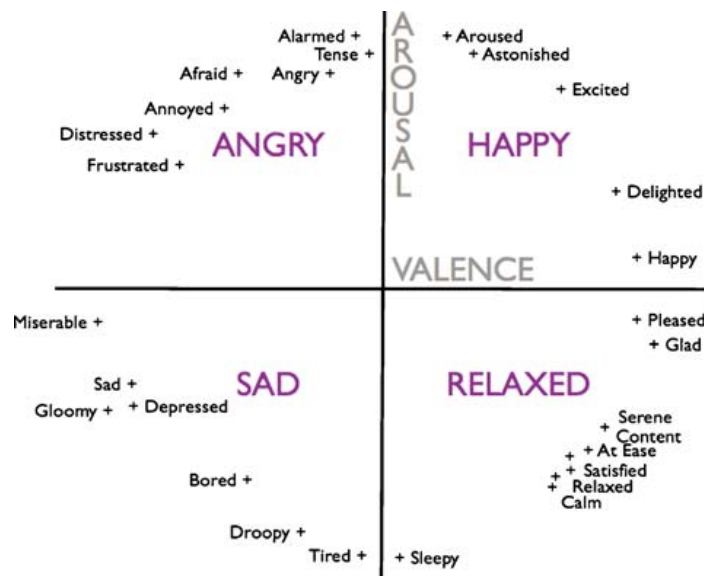
The Russell 2D model (called “circumplex model of affect”) is a reference widely accepted and cited in psychological studies on emotion. In this space, “happy” and “relaxed” have positive valence and, respectively, high and low arousal. “Angry” and “sad” have negative valence and, respectively, high and low arousal. As we do not want to be restricted to exclusive categories, we consider the problem as a binary classification task for each mood. One song can be “happy” or “not happy”, but also independently “angry” or “not angry” and so on.

The main idea of the present method is to exploit information extracted from both a social network and several experts validating the data. To do so, we have pre-selected the



**Fig. 1** Schema of the method employed to create the ground truth, validate it and design the music mood annotator

**Fig. 2** “Circumplex model of affect” (adapted from Russel [34])



tracks to be annotated using last.fm<sup>5</sup> tags (textual labels). Last.fm is a music recommendation website with a large community of users (30 million active users based in more than 200 countries) that is very active in associating tags with the music they listen to. These tags are then available to all users in the community. In Fig. 3, we show an example of a “tag cloud”, which is a visualization of the tags assigned to one song with the font size weighted by the popularity of the tag for this particular song.

In the example shown in Fig. 3, we can see that “happy” is present and quite highly weighted (which means that many people have used this tag to describe the song). In addition to “happy”, we also have “cheerful”, “joy”, “fun” and “upbeat”. To gather more data, we need to extend our query made to last.fm with more words related to mood. For the four chosen mood categories, we generated a set of related semantic words using Wordnet<sup>6</sup> and looked for the songs frequently tagged with these terms. For instance “joy”, “joyous”, “cheerful” and “happiness” are grouped under the “happy” category to generate a larger result set. We query the social network to acquire songs tagged with these words and apply a popularity threshold to select the best instances (we keep the songs that have been tagged by many users).

Note that the music for the “not” categories (like “not happy”) was evenly selected using both music tagged with antonyms and a random selection to create more diversity. Afterwards, we asked listeners to validate this selection. We considered a song to be valid if the tag was confirmed by, at least, one listener, as the pre-selection from last.fm granted that the song was likely to deserve that tag. We included this manual tag confirmation in order to exclude songs that could have received the tag by error, to express something else, or by a “following the majority” type of effect. The listeners were exposed to only 30 s of the songs to avoid changes in the mood as much as possible and to speed up the annotation process. Consequently, only these 30 s excerpts have been included in the final dataset. In total, 17 different evaluators participated and an average of 71% of the songs originally selected from last.fm was included in the training set. We observe that the “happy” and “relaxed” categories have a better validation rate than the “angry” and “sad” categories. This might be due to

<sup>5</sup> <http://www.last.fm>

<sup>6</sup> Wordnet is a large lexical database of English words with sets of synonyms <http://wordnet.princeton.edu/>



**Fig. 3** Tag cloud of the song “Here comes the sun” from the Beatles. The tags recognized as mood tags are underlined. The bigger the tag is, more people have used it to define that song

confusing terms in the tags used in the social networks for these latter categories or to a better agreement between people for “positive” emotions. These results indicate that the validation by experts is a necessary step to ensure the quality of the dataset. Otherwise, around 29% of errors, on average, would have been introduced. This method is relevant to pre-selecting a large number of tracks that potentially belong to one category.

At the end of the song selection process, the database was composed of 1,000 songs divided between the four categories of interest plus their complementary categories (“not happy”, “not sad”, “not angry” and “not relaxed”), i.e. 125 songs per category. The audio files were 30-second stereo clips at 44 khz in a 128 kbps mp3 format.

### 3.2 Audio feature extraction

In order to classify the music from audio content, we extracted a rich set of audio features based on temporal and spectral representations of the audio signal. For each excerpt, we merged the stereo channels into a mono mixture and its 200 ms frame-based extracted features were summarized with their component-wise statistics across the whole song. In Table 2, we present an overview of the extracted features by category.

For each excerpt we obtained a total of 200 feature statistics (minimum, maximum, mean, variance and derivatives), and we standardized each of them across the whole music collection values. In the next paragraphs, we describe some of the most relevant features for this mood classification task, with results and figures based on the training data.

#### 3.2.1 Mel frequency cepstral coefficients (MFCCs)

MFCCs [25] are widely used in audio analysis, and especially for speech research and music classification tasks. The method employed is to divide the signal into frames. For each frame, we take the logarithm of the amplitude spectrum. Then we divide it into bands and convert it to the perceptually-based Mel spectrum. Finally we take the discrete cosine transform (DCT).

**Table 2** Overview of the audio features extracted by category. See [12, 25, 31] for a detailed description of the features

Timbre	Bark bands, MFCCs, pitch salience, hfc, loudness, spectral: flatness, flux, rolloff, complexity, centroid, kurtosis, skewness, crest, decrease, spread
Tonal	dissonance, chords change rate, mode, key strength, tuning diatonic strength, tristimulus
Rhythm	bpm, bpm confidence, zero-crossing rate, silence rate, onset rate, danceability

The number of output coefficients of the DCT is variable, and is often set to 13, as we did in the present study. Intuitively, lower coefficients represent spectral envelope, while higher ones represent finer details of the spectrum. In Fig. 4, we show the mean values of the MFCCs for the “sad” and “not sad” categories. We note from Fig. 4 a difference in the shape of the MFCCs. This indicates a potential usefulness to discriminate between the two categories.

### 3.2.2 Bark bands

The Bark band algorithm computes the spectral energy contained in a given number of bands, which corresponds to an extrapolation of the Bark band scale [31, 38]. For each Bark band (27 in total) the power-spectrum is summed. In Fig. 5, we show an example of the Bark bands for the “sad” category.

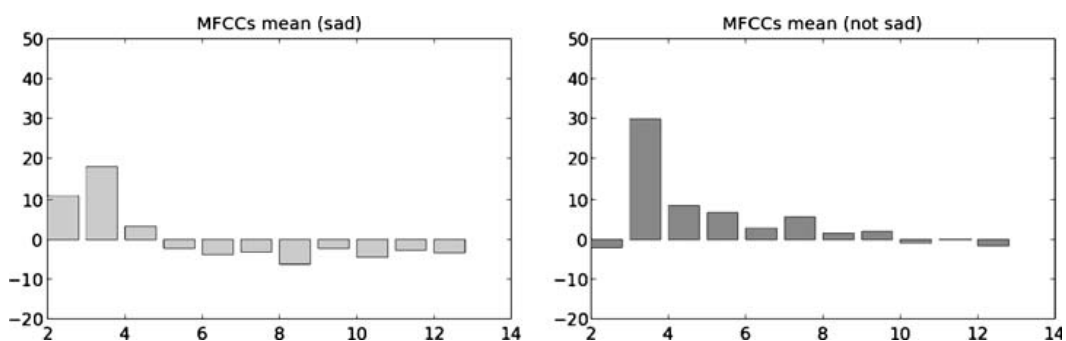
As with the MFCCs, the Bark bands appear to have quite different shapes for the two categories, indicating a probable utility for classification purposes.

### 3.2.3 Spectral complexity

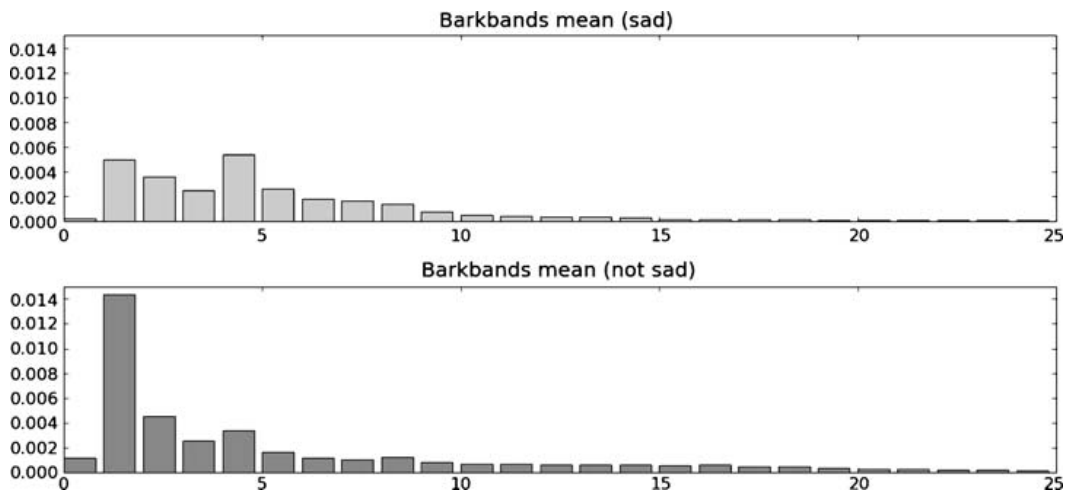
The spectral complexity descriptor is based on the number of peaks in the input spectrum. We apply peak detection on the spectrum (between 100 Hz and 5 KHz) and we count the number of peaks. This feature describes the complexity of the audio signal in terms of frequency components. In Figs. 6 and 7, we show the box-and-whisker plots of the spectral complexity descriptor’s standardized means for the “relaxed”, “not relaxed”, “happy” and “not happy” categories. These results are based on the entire training dataset. These plots illustrate the intuitive result that a relaxed song should be less “complex” than a non-relaxing song. Moreover, Fig. 7 tells us that happy songs are on average spectrally more complex.

### 3.2.4 Spectral centroid and skewness

The spectral centroid and skewness descriptors [31] (as well as spread, kurtosis, rolloff and decrease) are descriptions of the spectral shape. The spectral centroid is the barycenter of the spectrum, which considers the spectrum as a distribution of frequencies. The spectral skewness measures the asymmetry of the spectrum’s distribution around its mean value. The lower the value, the more energy exists on the right-hand side of the distribution, while more energy on the left side indicates a higher spectral skewness value. In Fig. 8 we show the spectral centroid’s box-and-whisker plot for “angry” and in Fig. 9 the spectral skewness for “sad”.



**Fig. 4** MFCC mean values for coefficients between 2 and 13 for the “sad” and “not sad” categories of our annotated dataset



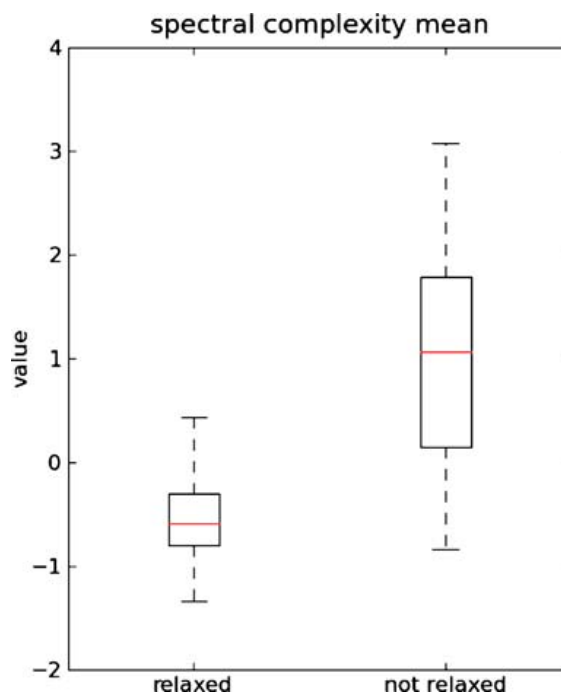
**Fig. 5** Bark band mean values for coefficients between 1 and 25 for the “sad” and “not sad” categories of our annotated dataset

Figure 8 shows a higher spectral centroid mean value for “angry” than “not angry”, which intuitively means more energy in higher frequencies. For the spectral skewness, the range of mean values for the “sad” instances is bigger than for the “not sad” ones. This probably means that there is a less specific value for the centroid. In any case, it seems to have on average a lower value for the “not sad” instances.

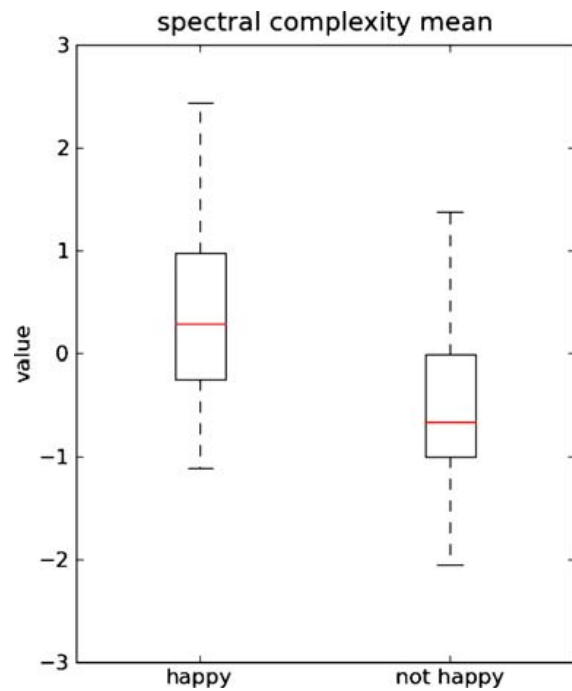
### 3.2.5 Dissonance

The dissonance feature (also known as “roughness” [35]) is defined by computing the peaks of the spectrum and measuring the spacing of these peaks. Consonant sounds have more

**Fig. 6** Box-and-whisker plot of the standardized spectral complexity mean feature for “relaxed” and “not relaxed”

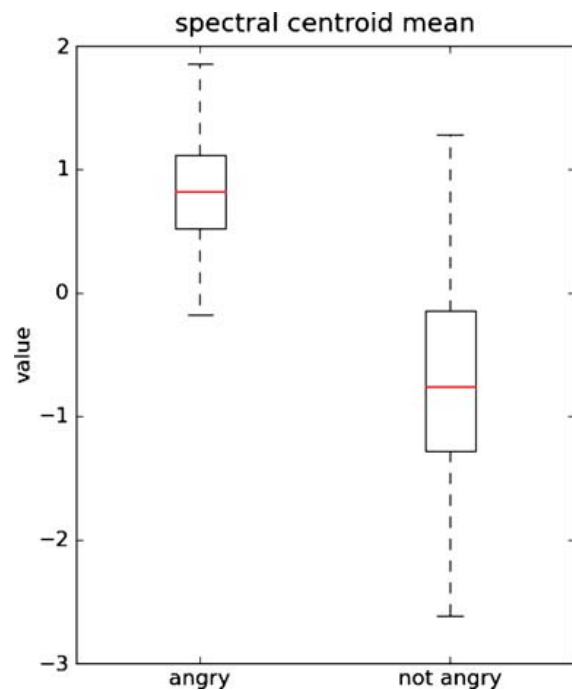


**Fig. 7** Box-and-whisker plot of the standardized spectral complexity mean feature for “happy” and “not happy”

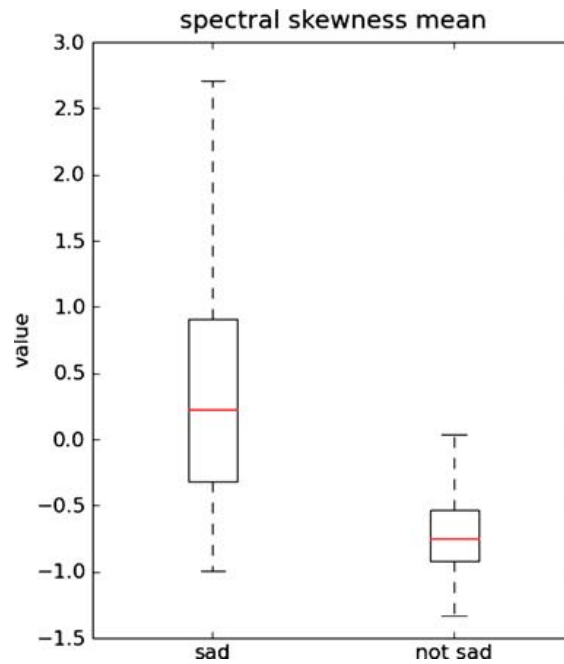


evenly spaced spectral peaks and, on the contrary, dissonant sounds have more sporadically spaced spectral peaks. In Figs. 10 and 11, we compare the dissonance distributions for the “relaxed” and “angry” categories. These figures show that “angry” is clearly more dissonant than “not angry”. Listening to the excerpts from the training data, we noticed many examples with distorted sounds like electric guitar in the “angry” category, which seems to be captured by this descriptor. This also relates to psychological studies stating that dissonant harmony may be associated with anger, excitement and unpleasantness [13, 44].

**Fig. 8** Box-and-whisker plot of the standardized spectral centroid mean for “angry” and “not angry”



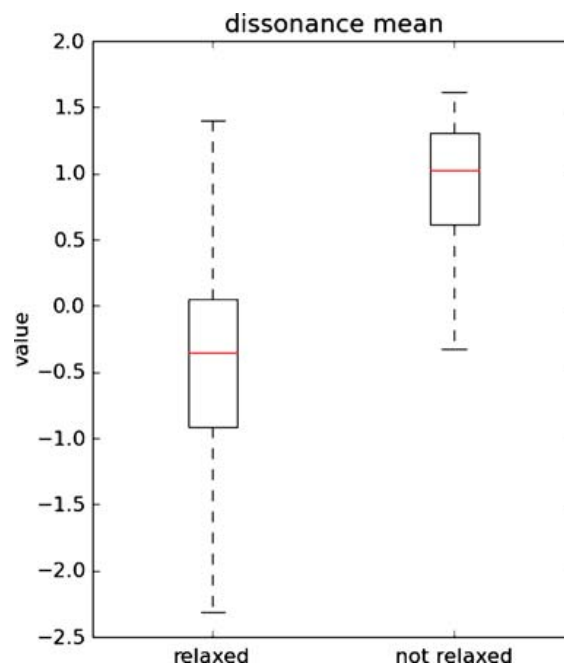
**Fig. 9** Box-and-whisker plot of the standardized spectral skewness mean for “sad” and “not sad”



### 3.2.6 Onset rate, chords change rate

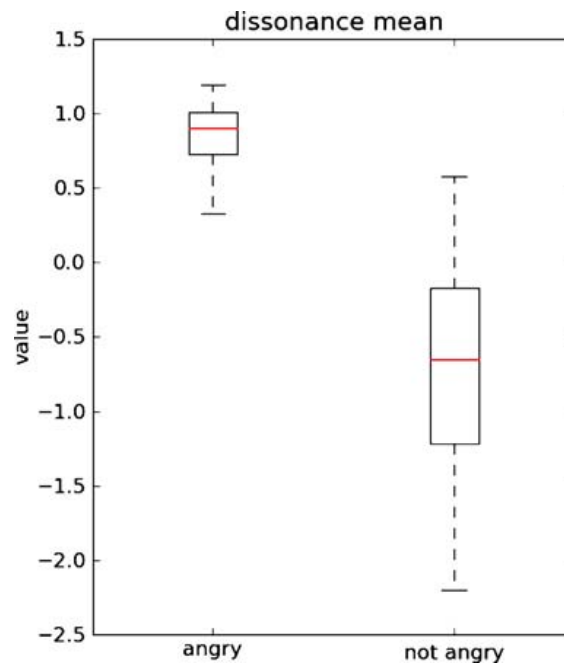
From psychological results, one important musical feature when expressing different mood types is rhythm (generally, faster means more arousal) [15]. The basic measure/element of rhythm is the onset, which is defined as an event in the music (any note, drum, etc.). The onset times are estimated by looking for peaks in the amplitude envelope. The onset rate is the number of onsets in one second. This gives us the number of events per second, which

**Fig. 10** Box-and-whisker plot of the standardized dissonance mean for the “relaxed” and “not relaxed” categories





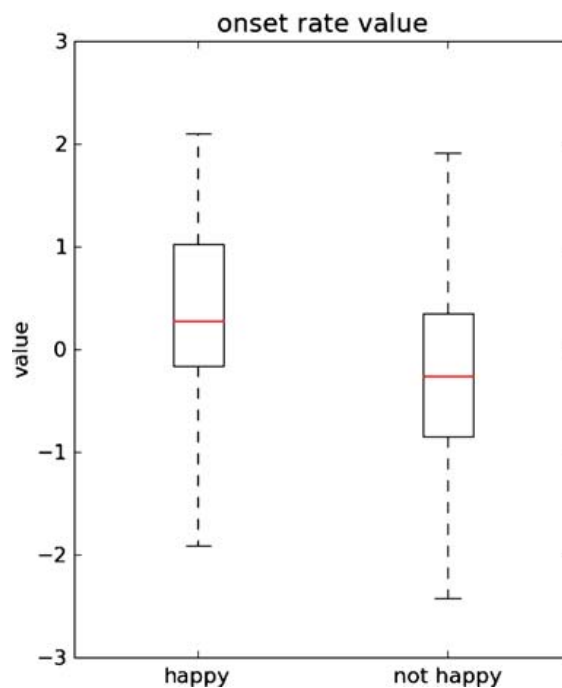
**Fig. 11** Box-and-whisker plot of the dissonance mean for the “angry” and “not angry” categories



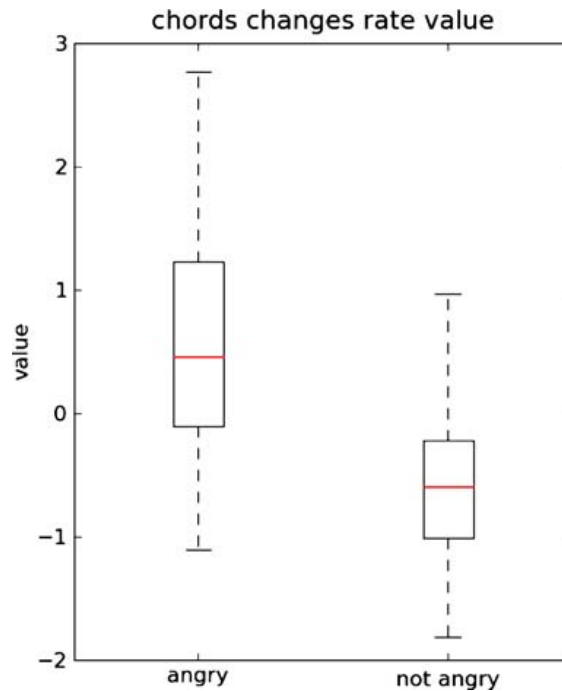
is related to a perception of the speed. The chords change rate is a rough estimator of the number of chords change per second.

In Fig. 12, we compare the onset rate values for the “happy” and “not happy” categories. It shows that “happy” songs have higher values for the onset rate, which confirms the psychological results that “happy” music is fast [15]. In Fig. 13, we look at the chords change rate, which is higher for “angry” than “not angry”. This is also a confirmation of the studies previously mentioned, associating higher arousal with faster music.

**Fig. 12** Box-and-whisker plot of the standardized onset rate value mean for the “happy” and “not happy” categories



**Fig. 13** Box-and-whisker plot of the chords change mean for the “angry” and “not angry” categories



### 3.2.7 Mode

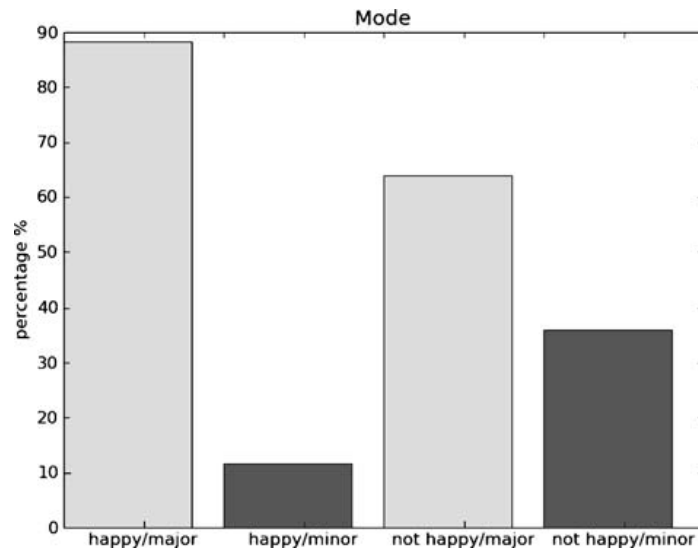
In Western music theory, there are two basic modes: major and minor. Each of them has different musical characteristics regarding the position of tones and semitones within their respective musical scales. Gómez [11] explains how to compute an estimation of the mode from raw audio data.

The signal is first pre-processed using the direct Fourier transform (DFT), filtering frequencies between 100 Hz and 5000 Hz and locating spectral peaks. The reference frequency (tuning frequency) is then estimated by analyzing the frequency deviation of the located spectral peaks. Next the Harmonic Pitch Class Profile (HPCP) feature is computed by mapping frequency and pitch class values (musical notes) using a logarithmic function [11]. The global HPCP vector is the average of the instantaneous values per frame, normalized to [0,1] to make it independent of dynamic changes. The resulting feature vector represents the average distribution of energy among the different musical notes. Finally, this vector is compared to minor and major reference key profiles based on music theory [17]. The profile with the highest correlation with the HPCP vector defines the mode.

In Fig. 14, we represent the percentages of estimated major and minor music in the “happy” and “not happy” categories. We note that there is more major music in the “happy” than in the “not happy” pieces. In music theory and psychological research, the link between valence (positivity) and the musical mode has already been demonstrated [15]. Still, having empirical data from an audio feature automatically extracted showing the same tendency is an interesting result. We note also that the proportion of major music is also high in the “not happy” category, which is related to the fact that the majority, 64%, of the whole dataset is estimated as major.

We have mentioned here some of the most relevant features showing their potential to individually discriminate between categories, however, we keep all the descriptors in our “bag-of-features”; those that are not obviously useful could be significant when combined

**Fig. 14** Bar plot of the estimated mode proportions (in percentage) for the “happy” and “not happy” categories



with others in a linear or non-linear way. To capture these relationships and build the model, we tried several kinds of classification algorithms.

### 3.3 Classification algorithms

Once the ground truth was created and the features extracted, we performed a series of tests with 8 different classifiers. We evaluated the classifiers using their implementations in Weka [46] with 10 runs of 10-fold cross-validation and parameter optimizations (See Table 3 for the mean accuracies). Next, we list the different classifiers we employed.

#### 3.3.1 Support Vector Machines (SVMs)

Support Vector Machines [3], is a widely used supervised learning classification algorithm. It is known to be efficient, robust and to give relatively good performance. Indeed, this classifier is widely used in MIR research. In the context of a two-class problem in  $n$  dimensions, the idea is to find the “best” hyperplane separating the points of

**Table 3** Mean classification accuracy with 10 runs of 10-fold cross-validation, for each category against its complementary. In bold is the highest accuracy for each category. The last column is the duration, in seconds, for a 10-fold cross-validation (computed on a 1.86 Ghz Intel Core Duo)

	Angry	Happy	Relaxed	Sad	Mean Accuracy	Duration 10 folds
SVM linear	95.79%	<b>84.57%</b>	90.68%	87.31%	89.58%	14 s
SVM poly	<b>98.17%</b>	84.48%	<b>91.43%</b>	<b>87.66%</b>	<b>90.44%</b>	24 s
SVM RBF	95.19%	84.47%	89.79%	87.52%	89.24%	17 s
SVM sigmoid	95.08%	84.52%	88.63%	87.31%	88.89%	17 s
J48	95.51%	80.02%	85.25%	85.87%	86.66%	5 s
Random Forest	96.31%	82.55%	89.47%	87.26%	88.90%	13 s
$k$ -NN	96.38%	80.89%	90.08%	85.48%	88.21%	4 s
Logistic Reg	94.46%	73.60%	82.54%	76.38%	81.75%	20 s
GMMs	96.99%	79.91%	91.13%	86.54%	88.64%	12 s

the two classes. This hyperplane can be of  $n-1$  dimensions and found in the original feature space, in the case that it is a linear classifier. Otherwise, it can be found in a transformed space of higher dimensionality using kernel methods (non-linear). The position of new observations compared to the hyperplane tells us in which class belongs the new input. For our evaluations, we tried different kernel methods: linear, polynomial, radial basis function (RBF) and sigmoid respectively called SVM linear, SVM poly, SVM RBF and SVM sigmoid, as shown in Table 3. To find the best parameters in each case we used the cross-validation method on the training data. For the linear SVM we looked for the best value for the cost  $C$  (penalty parameter), and for the others we applied a grid search to find the best values for the pair  $(C, \gamma)$  [3]. For  $C$ , we used the range  $[2^{-15}, 2^{15}]$  in 31 steps. For  $\gamma$ , we used the range  $[2^{15}, 2^3]$  in 19 steps. In the other cases than the linear SVM, once we have the best pair of values  $(C, \gamma)$ , we conduct a finer grid search on the neighborhood of these values. We note that from our data, the best parameter values highly depends on the category. Moreover, even if a RBF kernel is not always recommended for large feature sets compared to the size of the dataset [3], we achieved the best accuracy using this kernel for almost all categories. We used an implementation of the Support Vector Machines called libsvm [6].

### 3.3.2 Trees and random forest

The decision tree algorithm splits the training dataset into subsets based on a test attribute value. This process is repeated on each subset in a recursive manner (recursive partitioning). The random forest classifier uses several decision trees in order to improve the classification rate. We used an implementation of the C4.5 decision tree [33] (called J48 in Weka and in Table 3). To optimize the parameters of the decision tree, we performed a grid search on the two main parameters:  $C$  (the confidence factor used for pruning) from 0.1 to 0.5 in 10 steps and  $M$  (the minimum number of instances per leaf) from 2 to 20.

### 3.3.3 $k$ -Nearest Neighbor ( $k$ -NN)

For a new observation, the  $k$ -NN algorithm looks for a number  $k$  of the closest training samples to decide on the class to predict. The result relies mostly on the choice of distance function, which might not be trivial in our case, and also in the choice of  $k$ . We tested different values of  $k$  (between 1 and 20) with the Euclidean distance function.

### 3.3.4 Logistic regression

Logistic regression can predict the probability of occurrence of an event by fitting data to a logistic curve. It is a generalized linear model used for binomial regression. To optimize it, we varied the ridge value [21].

### 3.3.5 Gaussian Mixture Models (GMMs)

GMM is a linear combination of Gaussian probability distributions. This approach assumes that the likelihood of a feature vector can be expressed with a mixture of Gaussian distributions. GMMs are universal approximations of density, meaning that with enough Gaussians, any distribution can be estimated. In the training phase, the parameters of the Gaussian mixtures for each class are learnt using the Expectation-Maximization algorithm, which iteratively computes

maximum likelihood estimates [7]. The initial Gaussian parameters (means, covariance, and prior probabilities) used by the EM algorithm are generated via the k-means method [9].

### 3.4 Evaluation results

After independent parameter optimization for each classifier, the evaluation was made with 10 runs of 10 fold cross-validation. For comparison purposes, we show the mean accuracies obtained for each mood category and algorithm configuration separately. Each value in a cell represents the mean value of correctly classified data in the test set of each fold. Considering that each category is binary (for example, “angry” vs. “not angry”), the random classification accuracy is 50%.

The SVM algorithm with different kernels and parameters, depending on the category, achieved the best results. Consequently, we will choose the best configuration (SVM with polynomial kernel except for happy where we will use a linear SVM) for the integration in the final application.

The accuracies we obtained using audio-based classifiers are quite satisfying and even exceptional when looking at the “angry” category with 98%. All four categories reached classification accuracies above 80%, and two categories (“angry” and “relaxed”) peaked above 90%. Even though these results might seem surprisingly high, this is coherent with similar studies [37]. Also, the training examples were selected and validated only when they clearly belonged to the category or its complementary. This can bias the database and the model towards detecting very clear between-class distinctions.

### 3.5 Audio feature contribution

In this part, we evaluated the contribution of the audio features described in 3.2. In order to achieve this goal, we chose the best overall classifier for each category and we made 10 runs of 10-fold cross-validation with only one descriptor type statistic. We show in Table 4 the resulting mean accuracies for each configuration compared to the best accuracy obtained with all the features in the first row.

We observe that most of the descriptors give worst results for the “happy” category. This reflects also the results with all features, with a lower accuracy for “happy”. Moreover, some descriptors like the spectral centroid and the chords change rate do not seem to contribute positively for this category. We also note that the mode helps to discriminate

**Table 4** Mean classification accuracy with 10 runs of 10-fold cross-validation, for each category against its complementary with feature sets made of one descriptor statistic

	Angry	Happy	Relaxed	Sad
All features	98.17%	84.57%	91.43%	87.66%
MFCCs	89.47%	57.59%	83.87%	81.74%
Bark bands	90.98%	59.82%	87.10%	83.48%
Spectral complexity	95.86%	55.80%	88.71%	86.52%
Spectral centroid	89.47%	50.00%	85.48%	83.04%
Spectral skewness	77.44%	52.23%	73.38%	73.48%
Dissonance	91.73%	62.05%	82.66%	79.57%
Onset rate	52.63%	60.27%	63.31%	72.17%
Chords change rate	74.81%	50.00%	69.35%	68.26%
Mode	71.43%	64.73%	52.82%	52.08%

between “happy” and “not happy”, like seen in Fig. 14. It is also relevant for the “angry” category. However it does seem useful for “sad” against “not sad”. It is also worth noticing that if some individual descriptors can give relatively high accuracies, the global system combining all the features is significantly more accurate.

### 3.6 Audio encoding robustness

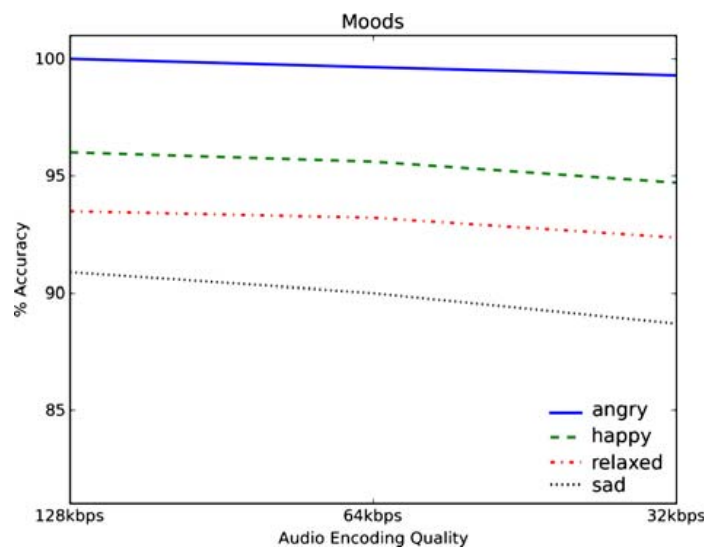
The cross-validation evaluation previously described gives relatively satisfying results in general. It allows us to select the best classifier with the appropriate parameters. However, since the goal is to integrate this model into a working platform, we have to test the stability and robustness of the mood classification with low quality encodings. Indeed it should be able to process musical content of different quality (commercial or user generated). The original encodings of the training set were mp3 at 128 kbps (kilobits per second). We generated two modified versions of the dataset, lowering the bit rate to 64 kbps and 32 kbps. In Fig. 15, we represent the accuracy degradation of the classifier trained with the entire dataset and tested on the same one with the previously mentioned low-rate encodings. We decided to train and test with full datasets, as this classifier model would be the one to be used in the final integrated version. Please note that the accuracies are different from Table 3 because in this case we are not performing cross-validation.

We observe degradation due to encoding at a lower bit rate. However, in all cases, this does not seem to have a strong impact. The degradation, in percentage, compared to the original version at 128 kbps is acceptable. For instance, we observe that for the “angry” category, at 32 kbps, only 0.7% of the dataset is no longer correctly classified as before. We also notice that the highest percentage of degradation is 3.6% obtained for the “relaxed” category (with 32 kbps). Even though there is a slight drop in the accuracy, the classification still gives satisfying results.

## 4 Integration in the PHAROS project

After explaining the method we used to build the ground truth, extract the features, select the best classification model and evaluate the results and robustness, we discuss here the integration of this technology in the PHAROS search engine framework.

**Fig. 15** Effect of the audio bit rate reduction on the accuracy (in percentage) for the entire dataset



#### 4.1 The PHAROS project

PHAROS<sup>7</sup> (Platform for searchIng of Audiovisual Resources across Online Spaces) is an Integrated Project funded by the European Union under the Information Society Technologies Programme (6th Framework Programme) - Strategic Objective 'Search Engines for Audiovisual Content'. PHAROS aims to advance audiovisual search from a point-solution search engine paradigm to an integrated search platform paradigm. One of the main goals of this project is to define a new generation of search engine, developing a scalable and open search framework that lets users search, explore, discover, and analyze contextually relevant data. Part of the core technology includes automatic annotation of content using integrated components of different kinds (visual classification, speech recognition, audio and music annotations, etc.). In our case, we implemented and integrated the automatic music mood annotation model previously described.

#### 4.2 Integration of the mood annotator

As a search engine, PHAROS uses automatic content annotation to index audiovisual content. However, there is a clear need to make the content analysis as efficient as possible (in terms of accuracy and time). To integrate mood annotation into the platform, we first created a fast implementation in C++ with proprietary code for audio feature extraction and dataset management together with the libsvm library for Support Vector Machines [6]. The SVMs were trained with full ground truth datasets and optimal parameters. Using a standard XML representation format defined in the project, we wrapped this implementation into a webservice, which could be accessed by other modules of the PHAROS platform. Furthermore, exploiting the probability output of the SVM algorithm, we provided a confidence value for each mood classifier. This added a floating point value that is used for ranking the results of a query by the annotation probability (for instance from the less to the most happy).

The resulting annotator extracts audio features and predicts the music's mood at a high speed (more than twice real-time), with the same performance level than what was presented in the previous section (using the same database). This annotator contributes to the overall system by allowing for a flexible and distributed usage. In our tests, using a cluster of 8 quad-core machines, we can annotate 1 million songs (using 30-seconds of each) in 10 days. The mood annotation is used to filter automatically the content according to the needs of users and helps them to find the content they are looking for. This integrated technology can lead to an extensive set of new tools to interact with music, enabling users to find new pieces that are similar to a given one, providing recommendations of new pieces, automatically organizing and visualizing music collections, creating playlists or personalizing radio streams. Indeed, the commercial success of large music catalogs nowadays is based on the possibility of allowing people to find the music they want to hear.

### 5 User evaluation

In the context of the PHAROS project, a user evaluation has been conducted. The main goal of these evaluations was to assess the usability of the PHAROS platform and in particular, the utility of several annotations.

---

<sup>7</sup> <http://www.pharos-audiovisual-search.eu>

## 5.1 Protocol

26 subjects participated in the evaluation. They were from the general public, between 18 and 40 years old (27 in average), all of them self-declared eager music listeners and last.fm users. The content processed and annotated for this user evaluation was made of 2092 30-second music videos. After a presentation of the functionalities on site, the users were then directly using an online installation of the system from their home. During 4 weeks, they could test it with some tasks they were asked to do every two days. The task related to our algorithm was to search for some music and to refine the query using a mood annotation. One query example could be to search for “music” and then refine with the mood annotation “relaxed”. They had to answer a questionnaire at the end of the study:

- “Do you find it interesting to use the mood annotation to refine a query for music?”
- “Do you find the “mood” annotation innovative?”
- “Does the use of the mood annotation correspond to your way of searching for audiovisual information?”

## 5.2 Results

As a general comment, there is difficulty for users to understand directly a content-based annotation. Some effort and thinking has to be done to make it intuitive and transparent. For instance what is “sad=0.58” (music annotated sad with a confidence of 0.58), is it really sad? Is it very sad? The confidence, or probability, value of one annotation is quite relative to other instances and most of all to the training set. This can be used for ranking results but might not be shown to the end-user directly. We should prefer nominal values like “very sad” or “not sad” for instance. Another important point seen when analyzing the comments from the users is the need to focus on precision. Especially in the context of a search engine, people will only concentrate on the first results and may not go to the second page. Instead, they are more likely to change their query. Several types of musical annotations were proposed to the user (genre, excitement, instrument, color, mode and key). From this list, mood was ranked as the second best in utility, just after musical genre (which is often given as metadata). Users had to rate on a scale from 0 to 10 their answer to several questions (0 would be “I strongly disagree” and 10 “I strongly agree”). We summarize here the answers to the questions related to the mood annotation:

- “Do you find it interesting to use the mood annotation to refine a query for music?” Users answered positively with a mean of 8.66, standard deviation of 1.85, showing a great interest to use this annotation.
- “Do you find the “mood” annotation innovative?” The mean of answers was also positive with 6.18 in average (standard deviation 3.81).
- “Does the use of the mood annotation correspond to your way of searching for audiovisual information?” Here users agreed with an average of 6.49 (standard deviation 3.47).

In all cases the mood annotation and its integration into the PHAROS platform was greatly accepted and highly considered by users. They also rated it as the most innovative musical annotation overall. In Fig. 16, we show a screenshot of the version of the PHAROS platform installed for the user evaluation. As an open framework, a PHAROS installation can be instantiated with different configurations, features and user interfaces. In this study



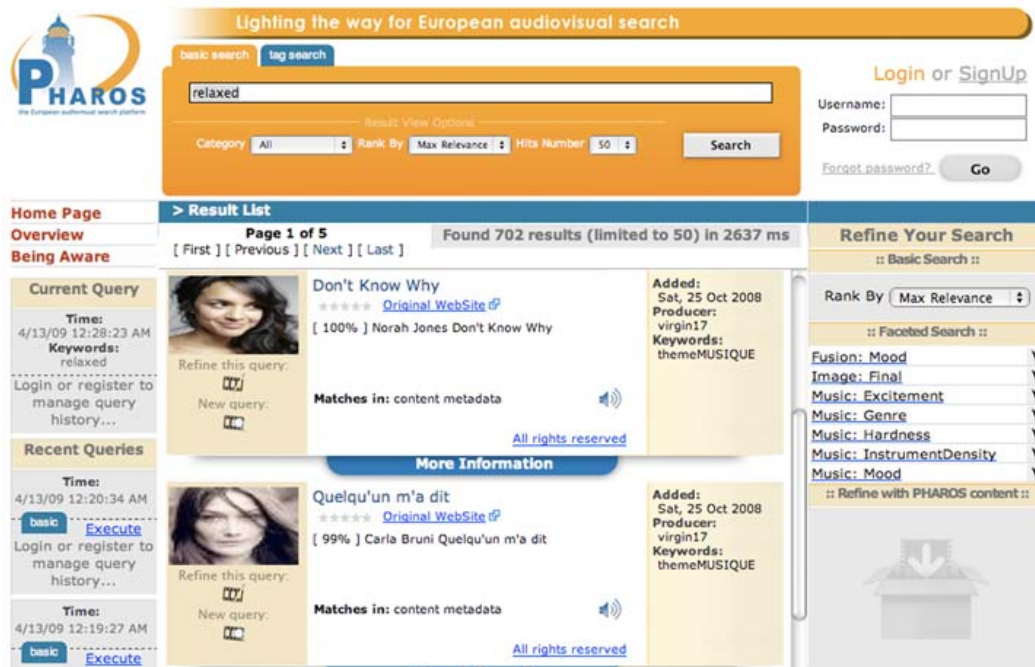


Fig. 16 Screenshot of the PHAROS interface used for the user evaluation

we used an instance created by taking advantage of Web Ratio<sup>8</sup> (an automatic tool to generate web interface applications). In this screenshot, the user is searching for “relaxed” music. They enter “relaxed” as a keyword and are browsing the musical results. The ones shown here were rated as “relaxed” (respectively 100% and 99%) thanks to the automatic music mood annotator we describe in this article.

## 6 Discussion and conclusion

We presented an approach for automatic music mood annotation introducing a procedure to exploit both the wisdom of crowds and the wisdom of the few. We detailed the list of audio features used and revealed some results using those most relevant. We reported the accuracies of optimized classifiers and tested the robustness of the system against low bit rate mp3 encodings. We explained how the technology was integrated in the PHAROS search engine and used it to query for, refine and rank music. We also mentioned the results from a user evaluation, showing a real value for the users in an information retrieval context. However, one may argue that this approach with 4 mood categories is simple when compared to the complexity of human perception. This is most likely true. Nevertheless, this is an important first step for this new type of annotation. So what could be done to improve it? First, we can add more categories. Although there might be a semantic overlap, it can be interesting to annotate music moods with a larger vocabulary, if we can still have high accuracies and add useful information (without increasing the noise for the user). Then, we can try to make better predictions by using a larger ground truth dataset or by designing new audio descriptors especially relevant for this task. Another option would be to generate analytical features [30], or to combine several classifiers to try to increase the

<sup>8</sup> <http://www.webratio.com>

accuracy of the system. We could also consider the use of other contextual information like metadata, tags, or text found on the Internet (from music blogs for instance). It has also been shown that lyrics can help to classify music by mood [20]. Indeed, multimodal techniques would allow us to capture more emotional data but also social and cultural information not contained in the raw audio signal. We should also focus on the user's needs to find the best way to use the technology. There is a clear need to make the annotation more understandable and transparent. Mood representations can be designed to be more usable than only textual labels. Finally, the mood annotation could be personalized, learning from the user's feedback and his/her perception of mood. This would add much value, although it might require more processing time per user, thus making the annotation less scalable. Nevertheless, it could dramatically enhance the user experience.

**Acknowledgments** We are very grateful to all the human annotators that helped to create our ground truth dataset. We also want to thank all the people contributing to the Music Technology Group (Universitat Pompeu Fabra, Barcelona) technologies and, in particular, Nicolas Wack, Eduard Aylon and Robert Toscano. We are also grateful to the entire MIREX team, specifically Stephen Downie and Xiao. We finally want to thank Michel Plu and Valérie Botherel from Orange Labs for the user evaluation data and Piero Fraternali, Alessandro Bozzon and Marco Brambilla from WebModels for the user interface. This research has been partially funded by the EU Project PHAROS IST-2006-045035.

## References

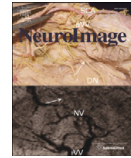
1. Andric A, Haus G (2006) Automatic playlist generation based on tracking user's listening habits. *Multimed Tools Appl* 29(2):127–151
2. Berenson ML, Goldstein M, Levine D (1983) *Intermediate statistical methods and applications: a computer package approach*. Prentice-Hall
3. Bigand E, Vieillard S, Madurell F, Marozeau J, Dacquet A (2005) Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts. *Cognition & Emotion* 19(8):1113–1139
4. Boser BE, Guyon IM, Vapnik VN (1992) A training algorithm for optimal margin classifiers. In *COLT '92: Proceedings of the fifth annual workshop on Computational learning theory*. ACM, New York, pp 144–152
5. Casey MA, Veltkamp R, Goto M, Leman M, Rhodes C, Slaney M (2008) Content-based music information retrieval: Current directions and future challenges. *Proc IEEE* 96(4):668–696
6. Chang CC, Lin CJ (2001) LIBSVM: a library for support vector machines, Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>
7. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the em algorithm. *J R Stat Soc, B* 39(1):1–38
8. Downie JS (2008) The music information retrieval evaluation exchange (2005–2007): a window into music information retrieval research. *Acoust Sci Technol* 29(4):247–255
9. Duda RO, Hart PE (1973) *Pattern classification and scene analysis*. Wiley, Somerset
10. Farnsworth PR (1954) A study of the Hevner adjective list. *J Aesthet Art Crit* 13(1):97–103
11. Gómez E (2006) Tonal description of music audio signals. PhD thesis, Universitat Pompeu Fabra
12. Gouyon F, Herrera P, Gómez E, Cano P, Bonada J, Loscos A, Amatriain X, Serra X (2008) *Content Processing of Music Audio Signals*, chapter 3, pages 83–160. Logos Verlag Berlin GmbH, Berlin
13. Hevner K (1936) Experimental studies of the elements of expression in music. *Am J Psychol* 58:246–268
14. Hu X, Downie JS, Laurier C, Bay M, Ehmann AF (2008) The 2007 MIREX audio mood classification task: Lessons learned. In *Proceedings of the 9th International Conference on Music Information Retrieval*, pp 462–467, Philadelphia, PA, USA, 2008
15. Juslin PN, Laukka P (2004) Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. *Journal of New Music Research*, 33(3)
16. Juslin PN, Västfjäll D (2008) Emotional responses to music: the need to consider underlying mechanisms. *Behavioral and Brain Sciences*, 31 (5)

17. Krumhansl CL (1997) An exploratory study of musical emotions and psychophysiology. *Can J Exp Psychol* 51(4):336–353
18. Laurier C, Herrera P (2007) Audio music mood classification using support vector machine. *Music Information Retrieval Evaluation eXchange (MIREX) extended abstract*
19. Laurier C, Herrera P (2009) Automatic detection of emotion in music: interaction with emotionally sensitive machines. *Handbook of Research on Synthetic Emotions and Sociable Robotics*. IGI Global, pp 9–32
20. Laurier C, Grivolla J, Herrera P (2008) Multimodal music mood classification using audio and lyrics. In *Proceedings of the International Conference on Machine Learning and Applications*. San Diego, CA, USA
21. Le Cessie S, Van Houwelingen JC (1992) Ridge estimators in logistic regression. *Appl Stat* 41(1):191–201
22. Li T, Ogihara M (2003) Detecting emotion in music. In *Proceedings of the 4th International Conference on Music Information Retrieval*, pages 239–240, Baltimore, MD, USA
23. Lidy T, Rauber A, Pertusa A, Iñesta JM (2007) MIREX 2007: combining audio and symbolic descriptors for music classification from audio. *MIREX 2007 — music information retrieval evaluation eXchange*, Vienna, Austria, September 23–27, 2007
24. Lindström E (1997) Impact of melodic structure on emotional expression. In *Proceedings of the Third Triennial ESCOM Conference*, pp 292–297
25. Logan B (2000) Mel frequency cepstral coefficients for music modeling. In *Proceeding of the 1st International Symposium on Music Information Retrieval*, Plymouth, MA, USA, 2000
26. Lu D, Liu L, Zhang H (2006) Automatic mood detection and tracking of music audio signals. *IEEE Trans Audio Speech Lang Process* 14(1):5–18
27. Mandel M, Ellis DP (2007) Labrosa’s audio music similarity and classification submissions. *MIREX 2007 — Music Information Retrieval Evaluation eXchange*, Vienna, Austria, September 23–27, 2007
28. Mandel M, Poliner GE, Ellis DP (2006) Support vector machine active learning for music retrieval. *Multimedia Systems*, 12(1)
29. Orio N (2006) Music retrieval: a tutorial and review. *Found Trends Inf Retr* 1(1):1–96
30. Pachet F, Roy P (2009) Analytical features: a knowledge-based approach to audio feature generation. *EURASIP Journal on Audio, Speech, and Music Processing* (1)
31. Peeters G (2004) A large set of audio features for sound description (similarity and classification) in the CUIDADO project. *Tech. rep.*, IRCAM
32. Peretz I, Gagnon L, Bouchard B (1998) Music and emotion: perceptual determinants, immediacy, and isolation after brain damage. *Cognition* 68(2):111–141
33. Quinlan RJ (1993) *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc, San Francisco
34. Russell JA (1980) A circumplex model of affect. *J Pers Soc Psychol* 39(6):1161–1178
35. Sethares WA (1998) *Tuning timbre spectrum scale*. Springer-Verlag
36. Shi YY, Zhu X, Kim HG, Eom KW (2006) A tempo feature via modulation spectrum analysis and its application to music emotion classification. In *Proceedings of the IEEE International Conference on Multimedia and Expo Toronto, Canada*, pp 1085–1088
37. Skowronek J, McKinney MF, van de Par S (2007) A demonstrator for automatic music mood estimation. In *Proceedings of the International Conference on Music Information Retrieval*, Vienna, Austria
38. Smith JO, Abel JS (1999) Bark and erb bilinear transforms. *IEEE Trans Speech Audio Process* 7(6):697–708
39. Sordo M, Laurier C, Celma O (2007) Annotating music collections: how content-based similarity helps to propagate labels. In *Proceedings of the 8th International Conference on Music Information Retrieval*, Vienna, Austria, pp 531–534
40. Thayer RE (1996) *The origin of everyday moods: managing energy, tension, and stress*. Oxford University Press, Oxford
41. Tzanetakis G (2007) Marsyas-0.2: a case study in implementing music information retrieval systems. In *Intelligent Music Information Systems*
42. Tzanetakis G, Cook P (2002) Musical genre classification of audio signals. *IEEE Trans Audio Speech Lang Process* 10(5):293–302
43. Vieillard S, Peretz I, Gosselin N, Khalifa S, Gagnon L, Bouchard B (2008) Happy, sad, scary and peaceful musical excerpts for research on emotions. *Cognition & Emotion* 22(4):720–752
44. Wedin L (1972) A Multidimensional study of perceptual-emotional qualities in music. *Scand J Psychol* 13(4):241–257
45. Wieczorkowska A, Synak P, Lewis R, Ras Z (2005) Extracting emotions from music data. In *Foundations of Intelligent Systems*, Springer-Verlag, pp 456–465
46. Witten IH, Frank E (1999) *Data mining: practical machine learning tools with Java implementations*. Morgan Kaufmann, San Francisco
47. Yang YH, Lin YC, Su YF, Chen HH (2008) A regression approach to music emotion recognition. *IEEE Trans Audio Speech Lang Process* 16(2):448–457

Koelsch, S., Skouras S., Fritz T., **Herrera, P.**, Bonhage, C., Küssner, M. B. & Jacobs, A.M. (2013). The roles of superficial amygdala and auditory cortex in music-evoked fear and joy. *NeuroImage*. 81(1), 49-60.

DOI: <https://doi.org/10.1016/j.neuroimage.2013.05.008>

ISSN: 1053-8119



## The roles of superficial amygdala and auditory cortex in music-evoked fear and joy



Stefan Koelsch<sup>a,b,c,\*</sup>, Stavros Skouras<sup>a,b,c,1</sup>, Thomas Fritz<sup>d</sup>, Perfecto Herrera<sup>e</sup>, Corinna Bonhage<sup>d</sup>, Mats B. Küssner<sup>f,g</sup>, Arthur M. Jacobs<sup>a,c</sup>

<sup>a</sup> Cluster Languages of Emotion, Freie Universität Berlin, Germany

<sup>b</sup> University of Sussex, Falmer, UK

<sup>c</sup> Dahlem Institute for Neuroimaging of Emotion, Berlin-Dahlem, Germany

<sup>d</sup> Max Planck Institute for Human Cognitive and Brain Science, Leipzig, Germany

<sup>e</sup> Universitat Pompeu Fabra, Barcelona, Spain

<sup>f</sup> King's College London, UK

<sup>g</sup> Goldsmiths, University of London, UK

### ARTICLE INFO

#### Article history:

Accepted 2 May 2013

Available online 17 May 2013

#### Keywords:

Emotion

Music

Fear

Joy

fMRI

Auditory cortex

Superficial amygdala

### ABSTRACT

This study investigates neural correlates of music-evoked fear and joy with fMRI. Studies on neural correlates of music-evoked fear are scant, and there are only a few studies on neural correlates of joy in general. Eighteen individuals listened to excerpts of fear-evoking, joy-evoking, as well as neutral music and rated their own emotional state in terms of valence, arousal, fear, and joy. Results show that BOLD signal intensity increased during joy, and decreased during fear (compared to the neutral condition) in bilateral auditory cortex (AC) and bilateral superficial amygdala (SF). In the right primary somatosensory cortex (area 3b) BOLD signals increased during exposure to fear-evoking music. While emotion-specific activity in AC increased with increasing duration of each trial, SF responded phasically in the beginning of the stimulus, and then SF activity declined. Psychophysiological Interaction (PPI) analysis revealed extensive emotion-specific functional connectivity of AC with insula, cingulate cortex, as well as with visual, and parietal attentional structures. These findings show that the auditory cortex functions as a central hub of an affective-attentional network that is more extensive than previously believed. PPI analyses also showed functional connectivity of SF with AC during the joy condition, taken to reflect that SF is sensitive to social signals with positive valence. During fear music, SF showed functional connectivity with visual cortex and area 7 of the superior parietal lobule, taken to reflect increased visual alertness and an involuntary shift of attention during the perception of auditory signals of danger.

© 2013 Elsevier Inc. All rights reserved.

### Introduction

Of all emotions, fear is the one that has been investigated most intensely in affective neuroscience over the last decades. However, there is scarcity of functional neuroimaging studies on fear with music, and neural correlates of music-evoked fear have thus remained elusive. This stands in gross contrast to a long musical tradition of using musical means to evoke fear in the listener. The earliest theoretical treatise on such means is the Affektenlehre (“theory of affects”) of the Baroque, which prescribed musical methods and figures for imitating, or portraying (and thus, according to the Affektenlehre, summoning) emotions, including fear (Mattheson, 1739/1999). Among countless well-known examples of fear-evoking (Western) music are Handel’s

Messiah (“And He Shall Purify”), Mozart’s *Idomeneo*, the thunderstorm portrayed in Beethoven’s sixth symphony, Berlioz’ *Sonje d’une nuit du Sabbat*, Herrmann’s music for *Psycho*, and Penderecki’s *Polymorphia*.

Nevertheless, only two previous functional neuroimaging studies have investigated brain responses to fear-evoking music. One of these studies explored how fear music can enhance feelings of fear evoked by images (Baumgartner et al., 2006), but that study did not present fear music alone, thus leaving open the question as to which activation patterns would be evoked by fearful music alone (i.e., without negative visual images). The other study investigated how music evoking fear or joy can change the perception of neutral film clips (Eldar et al., 2007). The latter study also investigated brain responses evoked by the fear music alone (without film clips), compared to a baseline condition, in selected regions of interest (amygdala, anterior hippocampal formation, prefrontal cortex, and auditory cortex). However, no effects of fear music were observed without film-clips (nor effects of joy or neutral music without film-clips), neither in the amygdala, nor in the hippocampus or the prefrontal cortex. In addition, a study by Lerner et al.

\* Corresponding author at: Freie Universität Berlin, Cluster Languages of Emotion, Habelschwerdter Allee 45, 14195 Berlin, Germany. Fax: +49 30 83852887.

E-mail address: [koelsch@cbs.mpg.de](mailto:koelsch@cbs.mpg.de) (S. Koelsch).

<sup>1</sup> The first two authors contributed equally to this work.

(2009) showed that listening to fear-evoking music with closed eyes (compared to listening with open eyes), evoked greater activation than open eyes in the amygdala/anterior hippocampal formation and anterior temporal poles (this effect of eyes open/closed was not observed when listening to neutral music). Main effects of fear compared to neutral music were not reported in that study. Finally, a recent study by Trost et al. (2012) reported brain activations due to music-evoked “tension” (characterized by feelings of high arousal and low valence), under which the authors also subsumed “feelings of anxiety and suspense induced by scary music” (brain activations included bilateral superior temporal gyrus, right parahippocampal gyrus, motor and premotor areas, cerebellum, right caudate nucleus, and precuneus). Notably, the concept of “tension” also includes emotional phenomena not related to fear, such as emotional reactions to unexpected musical events (Huron, 2006), and, therefore, Trost et al. (2012) argued that it is not clear whether the observed brain activations were due to fear responses, or to more general feelings of tension and unease. Thus, there are no functional neuroimaging data available that would allow us to draw conclusions about neural correlates of music-evoked fear.

With regard to lesion studies, Gosselin et al. (2005) showed impaired recognition of scary music in epileptic patients following unilateral medial temporal lobe excision (including the amygdala). In that study, both patients with left or right medial temporal lobe resections showed impaired recognition of scary, but not happy or sad, music. Corroborating this finding, data from a patient with bilateral damage restricted to the amygdala showed a selective impairment in the recognition of scary and sad music (Gosselin et al., 2007), indicating that the recognition of fear expressed by music involves the amygdala. These findings are reminiscent of findings reporting similar impairment for the recognition of fearful faces (reviewed in Peretz, 2010), suggesting that scary music and fearful faces are processed, at least in part, by common cerebral structures. Supporting this assumption, patients with unilateral anteromedial temporal lobe excision were found to be impaired in the recognition of both scary music and fearful faces (Gosselin et al., 2011), with results in both tasks being correlated. This suggested a multimodal representation of fear within the amygdala (although recognition of fearful faces was preserved in some patients, while their recognition of scary music was impaired). However, due to the size of the lesions in the reported studies, it remains unclear which nuclei of the amygdaloid complex played a role in the reported findings.

Functional neuroimaging studies on fear evoked by visual stimuli, recall/imagery, or auditory (but not musical) stimuli have also implicated the amygdaloid complex (LeDoux, 2000), in particular the basolateral amygdala (BL), as well as a range of functionally connected structures in fear responses (e.g., Phan et al., 2002). Such structures include the auditory cortex in auditory fear conditioning paradigms (LeDoux, 2000), as well as a large array of both cortical and subcortical structures, such as cingulate and insular cortex, hippocampus, parahippocampal cortex, orbitofrontal cortex, dorsolateral prefrontal cortex, striate (visual) cortex, basal ganglia, cerebellum, as well as brainstem regions such as the periaqueductal gray (Roy et al., 2009; Stein et al., 2007; Williams et al., 2006).

Based on the reported findings, we aimed to investigate the role of the amygdaloid complex and the auditory cortex, including their functional connections, for fear evoked by music. The cultural practice of using music to evoke fear makes music an important means to investigate neural circuits underlying fear (Eerola and Vuoskoski, 2011), in addition to the vast number of studies using visual stimuli to investigate neural correlates of fear. Besides fear stimuli, the present study also used joyful and neutral music. Joy was chosen as positive emotion because, on the one hand, both joy and fear are considered as “basic emotions” (Ekman, 1999), and both the expression of joy as well as of fear in Western music can be recognized universally (Fritz et al., 2009). On the other hand, other than, e.g. peaceful music (which is also perceived as positive, e.g. Vieillard et al., 2008), arousal levels evoked by joy music can well be matched with those evoked by fear

music. Similarly, musical and acoustical parameters such as tempo and pitch variation can well be matched between joy and fear music. Moreover, joyful music was chosen to replicate results of previous studies. Although only a few previous functional neuroimaging studies specifically used “happy” (Brattico et al., 2011; Brown et al., 2004; Mitterschiffthaler et al., 2007) or “joyful” (Koelsch et al., 2006; Mueller et al., 2011) music, these studies, along with other studies investigating musical frissons (Blood and Zatorre, 2001; Salimpoor et al., 2011), or music evoking emotional responses with positive valence and high arousal (Trost et al., 2012) indicate a number of relatively consistent features, namely stronger BOLD signal intensity (a) in the auditory cortex (Brattico et al., 2011; Koelsch et al., 2006; Mitterschiffthaler et al., 2007; Mueller et al., 2011; Trost et al., 2012), (b) the ventral striatum (Blood and Zatorre, 2001; Brown et al., 2004; Koelsch et al., 2006; Menon and Levitin, 2005; Mitterschiffthaler et al., 2007; Trost et al., 2012), (c) the anterior insula (Blood and Zatorre, 2001; Brown et al., 2004; Koelsch et al., 2006), and (d) the anterior cingulate cortex (Blood and Zatorre, 2001; Janata, 2009; Mitterschiffthaler et al., 2007). Moreover, (e) several studies on music-evoked emotions showed signal changes in the anterior hippocampal formation in response to stimuli with positive emotional valence (e.g., Blood and Zatorre, 2001; Mueller et al., 2011; Trost et al., 2012). Based on these findings, we hypothesized increased BOLD signals in response to joy stimuli (compared to neutral or fear stimuli) in the auditory cortex, ventral striatum, insula, ACC, and hippocampal formation.

Another aspect of our study was the investigation of the temporal dynamics of emotion across time. To our knowledge, only two previous functional neuroimaging studies have investigated the temporal dynamics of neural correlates of emotion (for habituation-effects across an experimental session see Mutschler et al., 2010). A study by Salimpoor et al. (2011) reported that BOLD signal intensity increased (a) in the dorsal striatum during the anticipation of a music-evoked frisson, and (b) in the ventral striatum during the experience of the frisson (notably, additional PET data showed that these signal increases were related to dopaminergic synaptic activity in these structures). Another study (Koelsch et al., 2006), in which stimuli of 60 s were split into two 30-second halves, showed that significant signal differences between pleasant and unpleasant music were most pronounced during the second half of the trials. The structures with such temporal dynamics of activation included the auditory cortex, inferior fronto-lateral areas (area 45 and the posterior part of the inferior frontal sulcus), anterior insula, the amygdaloid complex (probably basolateral amygdala), hippocampal formation, temporal poles, and parahippocampal cortex (a similar trend was observed in the ventral striatum).

Particular care was taken with regard to the acoustic parameters of our stimuli: numerous acoustical features of the stimuli were measured, which allowed us (1) to match joy, fear, and neutral stimuli with regard to numerous acoustical parameters (e.g., pitch variation, tempo, intensity, and spectral flux), and (2) to introduce acoustical factors that differed between conditions as regressors of no interest in the analysis of fMRI data. Provided that no crucial acoustical features were missed, this enabled us to investigate the role of the auditory cortex with regard to its emotion-specific interfacing with limbic/paralimbic structures. Previous work has implicated auditory association cortex (auditory parabelt), as well as its connections with the lateral amygdala, in fear conditioning (LeDoux, 2000). However, auditory parabelt regions project to numerous limbic/paralimbic structures (such as orbitofrontal cortex, insula, and cingulate cortex; e.g. Petrides and Pandya, 1988; Smiley et al., 2007; Yukie, 1995), and the role of these auditory projections for emotional processes, and thus the role that the auditory cortex plays for emotional processes, is largely unknown.

#### Summary of hypotheses

Motivated by the reported findings, we tested whether music-evoked fear, as compared to neutral or joy stimuli, would elicit signal changes in

the basolateral nucleus of the amygdaloid complex. For joy, as compared to neutral or fear, we expected stronger BOLD signal intensity in the ventral striatum, auditory cortex, hippocampal formation, insula, and cingulate cortex. Finally, to explore neural networks underlying joy and fear, we performed a Psychophysiological Interaction (PPI) analysis using the peak voxels indicated by the contrast analysis between conditions as seed voxels. More specifically, we were interested in emotion-specific functional connectivity between amygdaloid complex and auditory cortex, between auditory cortex and insula, as well as between auditory cortex and cingulate cortex.

## Materials and methods

### Participants

18 individuals (aged 20–31 years,  $M = 23.78$ ,  $SD = 3.54$ , 9 females) took part in the experiment. All participants had normal hearing (as assessed with standard pure tone audiometry) and were right-handed (according to self-report). None of the participants was a professional musician, nor a music student. Seven participants had no formal musical training, eight participants had once received music lessons (mean duration of formal training was 2.81 years,  $SD = 2.36$ , instruments were: flute, drums, piano, violin, guitar and melodica) but had not played their instruments for several years ( $M = 8.83$ ,  $SD = 7.52$ ), and three participants had learned a musical instrument that they were still playing (mean duration of formal training was 12.5 years,  $SD = 3.5$ , instruments were: guitar, violin, piano and electric bass). Exclusion criteria were left-handedness, professional musicianship, past diagnosis of a neurological or psychiatric disorder, a score of  $\geq 13$  on Beck's Depression Inventory (BDI; Beck et al., 1993), excessive consumption of alcohol or caffeine during the 24 h prior to testing, and poor sleep during the previous night. All subjects gave written informed consent. The study was conducted according to the Declaration of Helsinki and approved by the ethics committee of the School of Life Sciences and the Psychology Department of the University of Sussex.

### Stimuli and procedure

Musical stimuli were selected to evoke (a) feelings of joy, (b) feelings of fear, or (c) neither joy nor fear (henceforth referred to as neutral stimuli). There were  $n = 8$  stimuli per category (the complete list of joy and fear stimuli is provided in Supplementary Table S1). Joy stimuli had been used in previous studies (e.g., Fritz et al., 2009; Koelsch et al., 2010a, 2011; Mueller et al., 2011) and consisted of CD-recorded pieces from various epochs and styles (classical music, Irish jigs, jazz, reggae, South American and Balkan music). Fear-evoking musical stimuli were excerpts from soundtracks of suspense movies and video games. To increase the fear-evoking effect of the fear stimuli, their relatively high acoustic roughness (see also next paragraph) was further increased: from each fear excerpt, two copies were obtained and pitch-shifted, one copy was shifted one semitone higher, the other copy a tritone lower (see also Fritz et al., 2009; Koelsch et al., 2006). Then, all three versions of one excerpt (original pitch, one semitone higher, and a tritone lower) were rendered as a single wav-file (pitch-shift and rendering was performed using Ableton Live, version 8.0.4, Ableton AG, Berlin, Germany). Neutral stimuli were sequences of isochronous tones, for which the pitch classes were randomly selected from a pentatonic scale. These tone sequences were generated using the MIDI (musical instrument digital interface) toolbox for Matlab (Eerola and Toivainen, 2004). Importantly, for each joy–fear stimulus pair (see below), a neutral control stimulus was generated that matched joy and fear stimuli with regard to tempo, F0 range (i.e., range of the fundamental frequency), and instrumentation (using the two main instruments or instrument groups of the respective joy–fear pair). To create stimuli that sounded like musical compositions played with real instruments (similar to the joy and fear stimuli), the tones from the MIDI sequences were set to

trigger instrument samples from a high quality natural instrument library (X-Sample Chamber Ensemble, Winkler & Stahl GbR, Detmold, Germany) and from the Ableton Instrument library (Ableton AG, Berlin, Germany). Stimuli were then rendered as wav-files using Ableton Live. Using Praat (version 5.0.29; Boersma, 2002), all excerpts (joy, fear, and neutral) were edited so that they all had the same length (30 s), 1.5 fade-in/fade-out ramps, and the same RMS power.

Importantly, joy and fear stimuli were chosen such that each joyful excerpt had a fearful counterpart that matched with regard to tempo (beats per minute), mean fundamental frequency, variation of fundamental frequency, pitch centroid value, spectral complexity, and spectral flux. This was confirmed by an acoustic analysis of the stimuli using 'Essentia', an in-house library for extracting audio and music features from audio files (<http://mtg.upf.edu/technologies/essentia>). The Essentia software was also used to specify acoustical differences between stimuli with regard to other acoustical factors: 177 acoustical descriptors were extracted frame-by-frame (frame length = 21.5 ms, 50% overlap), averaged along the entire duration of the file, and then compared between conditions (joy, neutral, fear) using one-way ANOVAs. Bonferroni-corrected significance-level was  $0.05/177 = 0.00028$  (lowering this threshold for one-sided tests, i.e. 0.00056, did not change any of the results). The extracted features represent acoustic and musical features used in music information retrieval, i.e., different combinations of them are used for predictive models of musically relevant categorizations such as genre detection, instrument detection, key and mode detection, or emotional expression. Although these features have mostly been validated in machine-learning contexts (Huq et al., 2010; Kim et al., 2010; Laurier, 2011), it is possible that they also play a role for human auditory perception. In addition, many of the used parameters have been validated in perceptual experiments, such as features related to spectral complexity, F0, and F0 variations (Agrawal et al., 2012; Alluri et al., 2012; Coutinho and Dibben, 2012; Juslin and Laukka, 2003; Kumar et al., 2012), sensory dissonance (Coutinho and Dibben, 2012; Koelsch et al., 2006; Plomp and Levelt, 1965; Vassilakis and Kendall, 2010), spectral flux (Coutinho and Dibben, 2012; Menon et al., 2002), spectral centroid (Coutinho and Dibben, 2012), spectral crest (Laurier, 2011), temporal modulation frequencies (Kumar et al., 2012), key strength (Alluri et al., 2012; Krumhansl, 1990), and pulse clarity (Alluri et al., 2012). Significant effects of condition were indicated for the following acoustic factors (with  $F$ -values in parentheses, degrees of freedom: 2, 21): (a) Mean (72.3) and variance (13.8) of  $F0$  salience (this measure is highest for single tones, intermediate for chords, and lowest for noises; note that mean F0 and variance of F0 did not differ between joy, fear, and neutral stimuli). The mean F0 salience was highest for neutral, intermediate for joy, and lowest for fear stimuli ( $p < .0001$  in all pairwise comparisons). This reflects that both joy and fear (but not neutral) stimuli contained numerous harmonies, and that fear (but not joy) stimuli contained numerous percussive sounds, as well as hissing and whooshing noises. (b) Mean (41.3) and variance (28.0) of sensory dissonance. Sensory dissonance was lowest for neutral, intermediate for joy, and highest for fear stimuli. Mean sensory dissonance differed significantly between joy and neutral ( $p < .0001$ ), between fear and neutral ( $p < .0001$ ), and between joy and fear stimuli ( $p < .05$ ). (c) Mean chord strength (25.2) and key strength (14.7); these factors measure how strongly a sound resembles the sound of a chord, and how clearly the sounds of a stimulus can be attributed to a key. Chord strength was higher for joy compared to fear stimuli ( $p < .0001$ ), as well as for joy compared to neutral stimuli ( $p < .0006$ ), whereas fear and neutral stimuli did not differ significantly from each other. Key strength was higher for joy compared to fear stimuli ( $p < .0001$ ), and for neutral compared to fear stimuli ( $p = .01$ ); joy and neutral stimuli did not differ significantly from each other ( $p > .15$ ). (d) Mean (30.0) and variance (16.4) of spectral flux (a measure of spectral variation within sounds), mean (30.0) spectral crest (a measure of the inhomogeneity, or noisiness, of the spectrum) and mean (10.6) spectral complexity (which correlates with the amount of different timbres that are present

in a piece). Mean spectral flux, spectral crest, and spectral complexity were lowest for neutral stimuli (with significant differences between neutral and joy, as well as between neutral and fear stimuli,  $p < .05$  in each test), and did not differ significantly between joy and fear stimuli ( $p > .2$  in each test).

Prior to the MRI session, participants were presented with short (12 s) versions of each stimulus to obtain familiarity ratings: Participants rated their familiarity with each piece on a four-point scale (ranging from “To my knowledge I have never heard this piece before”, to “I know this piece, and I know who composed or performed it”). Participants were then trained on the rating procedure, using 12 s long excerpts of musical pieces that did not belong to the stimulus set used in the fMRI scanning session.

During the fMRI scanning session, stimuli were presented in a pseudo-random order so that no more than two stimuli of each stimulus category (joy, fear, neutral) followed each other. Participants were asked to listen to the musical stimuli with their eyes closed (see also Lerner et al., 2009). Each musical stimulus was followed by an interval of 2 s in which a beep tone of 350 Hz and 1 s duration signaled participants to open their eyes and to commence the rating procedure. During the rating procedure, participants indicated how they felt at the end of each excerpt with regard to valence (‘pleasantness’), ‘arousal’, ‘joy’ and ‘fear’. That is, participants provided ratings about how they felt, and not about which emotion each stimulus was supposed to express (Gabrielson and Juslin, 2003; Juslin and Västfjäll, 2008). Ratings were obtained with 6-point Likert scales (ranging from “not at all” to “very much”). The time interval for the rating procedure was 12 s and each rating period was followed by a 4 s rest period (during which participants closed their eyes again), amounting to a total length of 48 s per trial (see Fig. 1). The entire stimulus set was presented twice during the fMRI scanning session. Musical stimuli were presented using Presentation (version 13.0, Neurobehavioral systems, Albany, CA, USA) via MRI compatible headphones (under which participants wore earplugs). Instructions and rating screens were delivered through MRI compatible liquid crystal display goggles (Resonance Technology Inc., Northridge, CA,

USA) with integrated eye-tracker that allowed us to guarantee that participants opened and closed their eyes according to the instruction.

#### MR scanning

Scanning was performed with a 3 T Siemens Magnetom TrioTim. Prior to the functional MR measurements, a high-resolution ( $1 \times 1 \times 1$  mm) T1-weighted anatomical reference image was acquired from each participant using a rapid acquisition gradient echo (MP-RAGE) sequence. Continuous Echo Planar Imaging (EPI) was used with a TE of 30 ms and a TR of 2000 ms. Slice-acquisition was interleaved within the TR interval. The matrix acquired was  $64 \times 64$  voxels with a field of view of 192 mm, resulting in an in-plane resolution of 3 mm. Slice thickness was 3 mm with an interslice gap of 0.6 mm (37 slices, whole brain coverage). The acquisition window was tilted at an angle of  $30^\circ$  relative to the AC-PC line in order to minimize susceptibility artifacts in the orbitofrontal cortex (Deichmann et al., 2002, 2003; Weiskopf et al., 2007). Given the duration of our stimuli (30 s), a continuous scanning design was required to perform the PPI analysis (so that enough data points were available for meaningful correlation estimations, see below).

#### Data analysis

fMRI data were processed using LIPSIA 2.1 (Lohmann et al., 2001). Data were corrected for slicetime acquisition and normalized into MNI-space-registered images with isotropic voxels of 3 cubic millimeters. A temporal highpass filter with a cutoff frequency of 1/90 Hz was applied to remove low frequency drifts in the fMRI time series, and a spatial smoothing was performed using a 3D Gaussian kernel and a filter size of 6 mm FWHM.

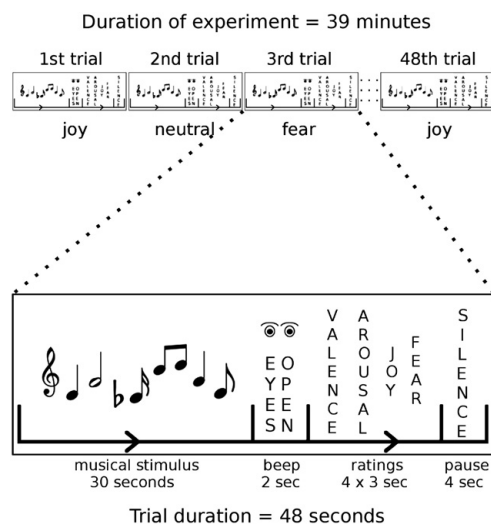
A mixed effects block design GLM analysis was employed (Friston et al., 2007). Valence ratings, arousal ratings, familiarity ratings, psychoacoustic parameters that differed significantly between conditions (see Stimuli and procedure), and realignment parameters were included in the design matrix as covariates of no interest (Johnstone et al., 2006). Then, one-sample  $t$ -tests were calculated voxel-wise for the contrast between fear vs. joy, and corrected for multiple comparisons by the use of cluster-size and cluster-value thresholds obtained by Monte Carlo simulations with a significance level of  $p < 0.05$  (Lohmann et al., 2008). The significant clusters identified in this analysis were used as regions of interest (ROIs) to compare the average signal intensity (averaged across all voxels in each cluster) within those clusters between fear and neutral, as well as between joy and neutral. In addition, to explore the temporal nature of the significant differences in activity between fear and joy, for each peak voxel of each significant cluster, the timecourse of activity was determined by computing the voxel intensity separately for each scan (i.e., with a temporal resolution of 2 s) and for each condition.

#### Temporal interaction analysis

To investigate possible interactions between emotion and time we split the data from each trial into first half (seconds 1 to 15) and second half (seconds 16 to 30), and calculated a statistical parametric map based on the interaction between emotion (two levels: joy, fear) and time (two levels: first half, second half). A first-level interaction contrast was calculated for each subject, and the contrast images were then used for voxel-wise one-sample  $t$ -tests at the second level (corrected for multiple comparisons by the use of cluster-size and cluster-value thresholds obtained by Monte Carlo simulations with a significance level of  $p < .05$ ) to identify clusters of voxels for which the emotion  $\times$  time interaction was significantly different from zero.

#### PPI analysis

The timecourses of activity at the peak voxels identified in the contrast joy vs. fear, averaged together with the timecourses from adjacent voxels, were used as seeds for Psychophysiological Interaction



**Fig. 1.** Experimental design. In each trial, a music stimulus was presented for 30 s. Music stimuli were pseudorandomly either a joy, a fear, or a neutral stimulus. Participants listened to the music with their eyes closed. Then, a beep tone signaled to open the eyes and to commence the rating procedure. Four ratings (felt valence, arousal, joy, and fear) were obtained in 12 s, followed by a 4 s pause (during which participants closed their eyes again). Trial duration was 48 s, the experiment comprised of 48 trials.



(PPI) analyses to identify target regions for which the covariation of activity between seed and target regions was significantly different between experimental conditions. At the first level, contrasts were calculated for each subject based on the interaction term between emotion (joy vs. fear) and each seed voxel's timecourse of activity (Friston et al., 1997). For each seed voxel, the contrast images from all subjects were used in voxel-wise one-sample *t*-tests at the second level (corrected for multiple comparisons by the use of cluster-size and cluster-value thresholds obtained by Monte Carlo simulations with a significance level of  $p < .05$ ) to identify clusters of voxels for which the psychophysiological interaction effect was significant.

## Results

### Behavioral data

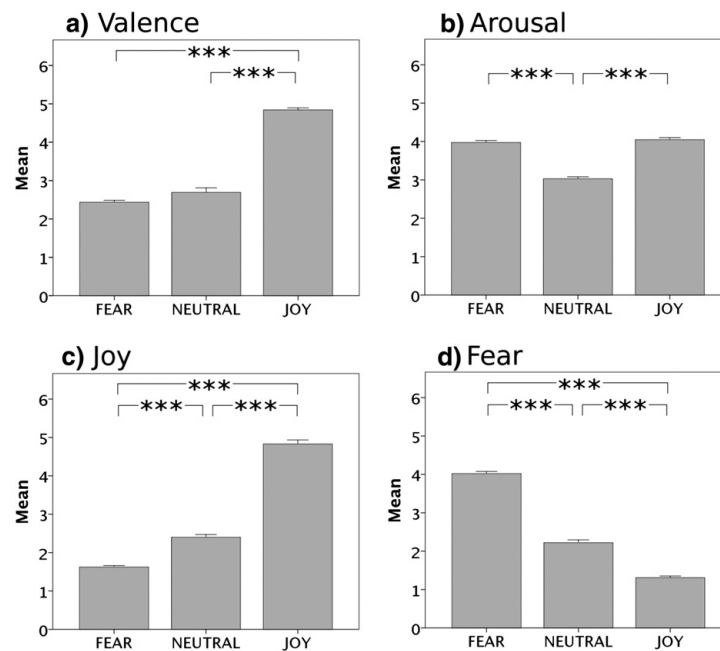
Behavioral data are summarized in Fig. 2 and Table 1. Valence (pleasantness) ratings were lower for fear than for joy stimuli ( $t(15) = 42.29, p < 0.0001$ ), higher for joy than for neutral stimuli ( $t(15) = 16.10, p < 0.0001$ ), and did not differ significantly between neutral and fear stimuli ( $t(15) = -1.94, p = .072$ ). Arousal ratings were higher for fear than for neutral stimuli ( $t(15) = 11.84, p < 0.0001$ ), higher for joy than for neutral stimuli ( $t(15) = 12.26, p < 0.0001$ ), and did not differ between joy and fear stimuli ( $t(15) = .94, p = .36$ ). Joy ratings were lowest for fear stimuli, and highest for joy stimuli, with ratings for neutral stimuli being in between. Joy ratings differed significantly between fear and neutral stimuli ( $t(15) = 9.03, p < 0.0001$ ), fear and joy stimuli ( $t(15) = 32.32, p < 0.0001$ ), and between joy and neutral stimuli ( $t(15) = 16.73, p < 0.0001$ ). Correspondingly, fear ratings were highest for fear stimuli, lowest for joy stimuli, with ratings for neutral stimuli

being in between. Although the degree of experienced fear was relatively moderate (4.02 on a scale from 1 to 6), fear ratings differed significantly between fear and neutral stimuli ( $t(15) = 17.71, p < 0.0001$ ), fear and joy stimuli ( $t(15) = 33.16, p < 0.0001$ ), and between joy and neutral stimuli ( $t(15) = 9.93, p < 0.0001$ ). Average familiarity ratings were highest for joy stimuli, lowest for neutral stimuli, with ratings for fear stimuli being in between. Familiarity ratings differed significantly between joy and fear stimuli ( $t(7) = 3.659, p < 0.05$ ), fear and neutral stimuli ( $t(7) = 4.41, p < 0.01$ ), and between joy and neutral stimuli ( $t(7) = 5.06, p < 0.0005$ ). Due to the differences in the behavioral ratings between stimulus categories with regard to valence, arousal, and familiarity, each participant's valence, arousal, and familiarity ratings were used in the fMRI data analysis as regressors of no interest (see Data analysis). Therefore, these variables (valence, arousal, and familiarity) did not contribute to the fMRI results presented in the following.

### fMRI data

#### GLM analysis

The statistical parametric maps (SPMs) of the contrast *joy > fear* (corrected for multiple comparisons,  $p < .05$ ) revealed significant BOLD signal differences in the auditory cortex (AC) bilaterally, and in the superficial amygdala (SF) bilaterally (see also Table 2 and Fig. 3a). The activation of the AC covered auditory core, belt, and parabelt regions bilaterally. The voxels with maximum *z*-values were located along Heschl's gyrus (HG), with the peak voxel in the left AC being located on the postero-lateral rim of HG (30% TE 1.2 according to Morosan et al., 2001), and the peak voxel in the right AC being located more medially on HG (90% TE 1.0 according to Morosan et al., 2001). In both left and right amygdala, the peak voxel was located in SF (left: 80% probability,



**Fig. 2.** Behavioral ratings of participants on the four emotion scales used in the present study: (a) valence, (b) arousal, (c) joy, and (d) fear. Range of scales was 1 to 6. Ratings are depicted separately for each stimulus category (fear, neutral, joy). Note that joy stimuli were rated as more pleasant than fear and neutral stimuli (valence/pleasantness ratings for fear and neutral stimuli did not differ from each other). Also note that arousal ratings of joy and fear stimuli did not differ from each other, and that both joy and fear stimuli were rated as more arousing than neutral stimuli.

**Table 1**

Descriptive statistics of behavioral data (mean, with standard deviation in parentheses). Range of valence, arousal, joy, and fear scales was 1 to 6, range of the familiarity scale was 1 to 4. For statistical tests see main text.

	Fear	Neutral	Joy
Valence	2.43 (0.20)	2.69 (0.48)	4.84 (0.21)
Arousal	3.97 (0.21)	3.03 (0.22)	4.05 (0.21)
Joyfulness	1.62 (0.16)	2.40 (0.29)	4.83 (0.42)
Fearfulness	4.02 (0.23)	2.22 (0.30)	1.31 (0.17)
Familiarity	1.44 (0.11)	1.17 (0.10)	2.01 (0.42)

right: 90% probability according to the cytoarchitectonic probability map by Amunts et al., 2005). The signal differences in SF extended bilaterally into the hippocampal-amygdaloid transition area (HATA, Amunts et al., 2005). The opposite contrast (*fear > joy*) showed signal differences in the anterior bank of the right postcentral gyrus (area 3b of the primary somatosensory cortex, S1, the peak voxel was located with 80% probability in this area according to Geyer et al., 1999). Contrasts with the neutral condition did not yield any additional activations (see also Table 2 and next section), except activations in the visual cortex for both *joy > neutral* (left V1, MNI-coordinate:  $-1, -82, -5$ ; left V4:  $-33, -82, -14$ ; right V2:  $32, -99, 3$ ) and *fear > neutral* (left V2, MNI-coordinate:  $-8, -95, 25$ ; right V2:  $23, -93, 26$ ).

#### ROI analysis

To specify whether the observed differences between fear and joy were due to signal increase or decrease compared to the neutral control condition, ROI analyses were conducted for the significant clusters identified in the GLM analysis (AC, SF, S1), comparing the mean signal intensity of the voxels in each cluster between fear and neutral, as well as between joy and neutral. Results of these analyses (corrected for multiple comparisons,  $p < .05$ ) showed that, compared to the neutral condition, there was stronger signal intensity during joy and weaker signal intensity during fear in the AC bilaterally as well as in the left SF (see also Table 2). In the right SF, signal intensity was weaker during fear compared to neutral (with no difference between joy and neutral). In the right S1, signal intensity was stronger during fear compared to neutral (joy and neutral did not differ from each other).

#### Timelines

To explore the temporal dynamics of the observed differences, the signal intensity of the peak voxel of each significant cluster (AC, SF, S1) was computed separately for each scan (i.e., with a temporal resolution of 2 s) in each condition. These timelines are shown in Fig. 4. In the AC, the auditory stimuli evoked a signal increase (in all conditions), with the signal intensity being generally highest for joy, lowest for fear, and intermediate for neutral (see next paragraph for statistical analysis). The most pronounced differences between conditions emerged at, and after around 10 s after stimulus onset. In SF, joy stimuli evoked a signal increase bilaterally, while fear stimuli evoked a signal increase only in the right SF. In the left SF, differences in signal intensity between fear and joy were particularly strong during the first half of the stimuli

**Table 2**

Results of General Linear Model (GLM) contrasts, corrected for multiple comparisons ( $p < .05$ ): (a) *joy > fear*, (b) *fear > joy*. The two outermost right columns provide the  $p$ -values for comparisons involving the neutral condition within the significant clusters identified in the GLM analysis (region of interest analysis). The diamonds in the outermost right column indicate that differences between fear and neutral were due to higher signal intensity during neutral than during fear. Abbreviations: ROI: region of interest; l: left; r: right; n.s.: not significant.

	MNI coordinate	Cluster size (mm <sup>3</sup> )	z-Value: max (mean)	p-Value ROI: joy vs. neutral	p-Value ROI: fear vs. neutral
(a) <i>joy &gt; fear</i>					
l Heschl's gyrus	$-56 -14 7$	16,038	6.36 (3.76)	.0002	.0001 <sup>◇</sup>
r Heschl's gyrus	$50 -16 8$	13,176	5.55 (3.72)	.0006	.0007 <sup>◇</sup>
l superficial amygdala	$-17 -7 -15$	486	4.32 (3.36)	.02	.009 <sup>◇</sup>
r superficial amygdala	$22 -6 -13$	324	3.40 (3.09)	n.s.	.03 <sup>◇</sup>
(b) <i>fear &gt; joy</i>					
r postcentral gyrus (area 3b)	$52 -13 36$	297	$-3.50 (-3.11)$	n.s.	.0001

(and a similar trend is observable in the right SF). Differences between conditions emerged several seconds after stimulus onset, were most pronounced at around 10 s, and vanished towards the end of the stimuli (see next paragraph for statistical analysis). In the right S1, all conditions evoked an initial signal decrease, followed by a signal increase (which was strongest for fear stimuli), and a decline of signal intensity towards the end of the stimuli.

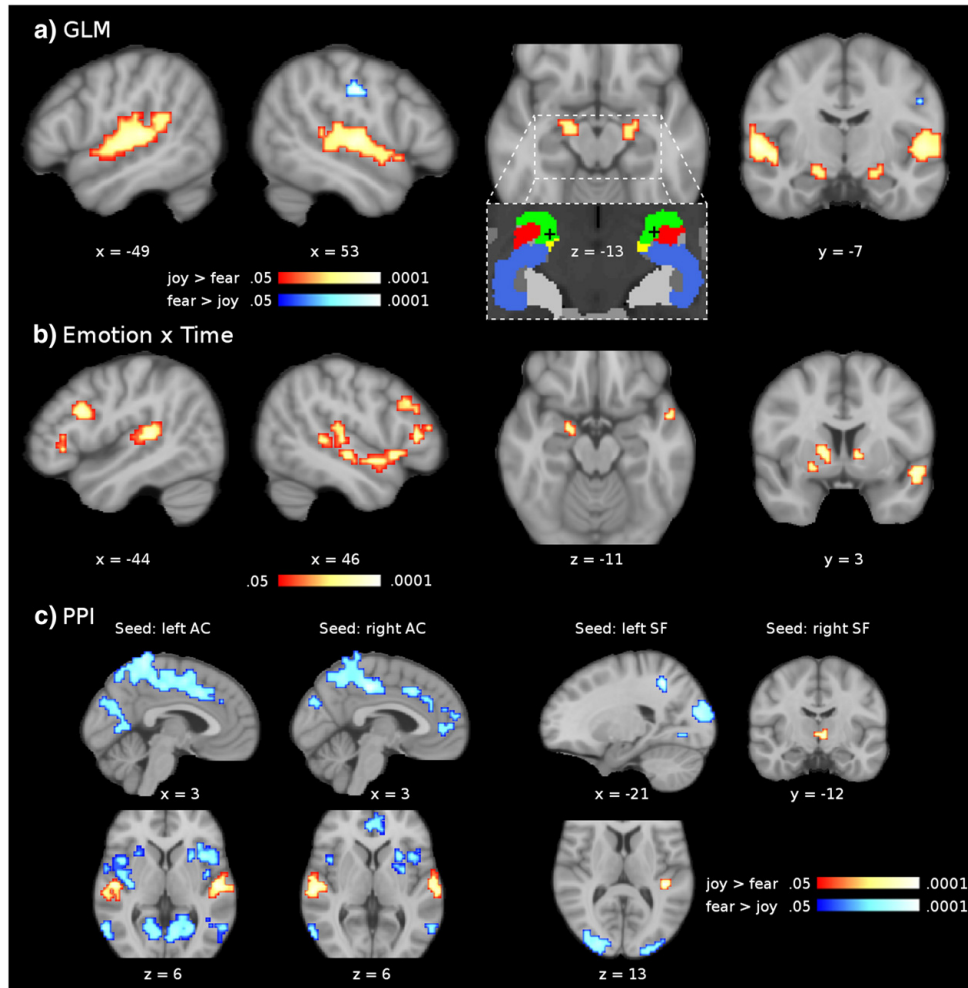
#### Temporal interaction analysis

To statistically test the temporal dynamics observed in the timelines, and to further explore the temporal dynamics of differences between conditions in other structures (see Introduction), a temporal interaction analysis was computed with factors emotion (two levels: joy and fear) and time (two levels: first half and second half of each stimulus, see Materials and methods). Results (corrected for multiple comparisons,  $p < .05$ ) are listed in Table 3 and summarized in Fig. 3b. Significant interactions were observed in the AC bilaterally, and in the left SF. This confirms the observations based on the timelines that differences in the AC were more pronounced during the second half, and in the SF during the first half of trials. Moreover, according to the hypotheses (see Introduction), significant emotion  $\times$  time interactions were observed bilaterally (a) in the posterior portion of the inferior frontal sulcus (IFS), (b) the anterior part of Broca's area (BA 45/46), and (c) in the ventral pallidum/ventral striatum (see also Fig. 3b). These interactions were due to more pronounced differences between conditions in the second compared to the first half. No interactions were observed in the hippocampus, parahippocampal gyrus, temporal poles, nor in the Rolandic operculum.

#### PPI analysis

Finally, we conducted a Psychophysiological Interaction (PPI) analysis (for details see Materials and methods). Seed regions were the peak voxels (as well as the directly adjacent voxels) identified in the GLM analysis in the direct contrast between fear and joy stimuli. Results of this analysis (corrected for multiple comparisons,  $p < 0.05$ ), are listed in Table 4 and summarized in Fig. 3c.

Both left and right AC showed stronger functional connectivity during joy (compared to fear) with both ipsilateral and contralateral AC. In specific, the left posterior-lateral auditory belt showed stronger functional connectivity during joy with both left and right primary auditory cortex (left: 80%, right: 100% probability for TE 1.0 according to Morosan et al., 2001), as well as with lateral auditory belt-regions of both hemispheres. The right auditory core region showed stronger functional connectivity during joy with lateral auditory belt regions of both hemispheres (TE 2 according to Morosan et al., 2001, no probabilistic maps are available for this region). During fear (compared to joy), both left and right AC showed stronger functional connectivity with the cuneus (areas 17 and 18), the median wall of the precuneus (areas 5 and 7), and almost the entire cingulate sulcus (CS), from the pre-genual CS to the ascending branch of the (posterior) CS. Moreover, both left and right AC showed stronger functional connectivity during fear with the anterior insula bilaterally, and the left (but not



**Fig. 3.** fMRI results (all corrected for multiple comparisons,  $p < .05$ ). (a) shows the statistical parametric maps (SPMs) of the direct contrast between joy and fear stimuli, red: joy > fear, blue: fear > joy. The SPMs show stronger BOLD signals during joy (compared to fear) in the auditory cortex (AC), and the SF bilaterally. Stronger BOLD signals during fear (compared to joy) were yielded in area 3b of the primary sensory cortex. The inset shows the coordinates of the peak voxels in the SF (indicated by the black crosses) projected on the cytoarchitectonic probability map according to Eickhoff et al. (2005); green: superficial amygdala, red: basolateral amygdala, yellow: hippocampal-amygdaloid transition area, blue: hippocampus (cornu ammonis). (b) shows the interaction contrast between emotion (joy vs. fear) and time (1st half of each trial vs. 2nd half of each trial). Significant interactions were indicated in the auditory cortex bilaterally, the left SF, left area 45 (pars triangularis of the inferior frontal gyrus), inferior frontal sulcus, and ventral pallidum/ventral striatum. (c) shows results of the Psychophysiological Interaction Analysis (PPI) for the regions that significantly differed in the SPM contrast between fear and joy, seed voxels were located in left AC (Heschl's gyrus), right AC (Heschl's gyrus), left SF, and right SF. Red/yellow colors indicate regions that exhibited stronger functional connectivity with the seed regions during the joy than during the fear condition. Blue colors indicate regions that exhibited stronger functional connectivity with the seed regions during the fear than during the joy condition.

the right) AC showed stronger functional connectivity during fear with the fundus of the central sulcus and the anterior bank of postcentral gyrus (areas 3a & b of S1).

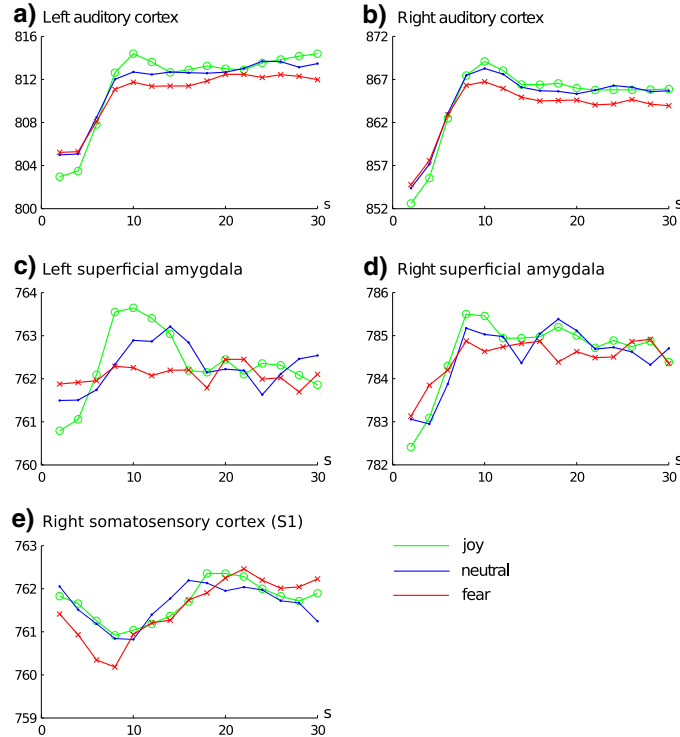
The left SF showed stronger functional connectivity during joy (compared to fear) with right posterior HG (posterior auditory core and belt regions, 80% TE 1.1 according to Morosan et al., 2001). During fear (compared to joy), the left SF showed stronger functional connectivity with cuneus (V1–V4), and area 7a of the superior parietal lobule (precuneus) bilaterally (left: 70%, right: 40% probability according to Scheperjans et al., 2008). The right SF showed stronger functional

connectivity during joy (compared to fear) with the mediodorsal nucleus of the thalamus (43% th-temporal according to Eickhoff et al., 2005). The S1 region did not show any significant PPI results in our data.

## Discussion

### Summary of results

The contrast analysis showed that BOLD signals in the auditory cortex (AC) bilaterally were strongest during joy, weakest during fear, with



**Fig. 4.** Timelines depicting average BOLD signal intensity in the regions that significantly differed in the GLM contrast joy vs. fear. The ordinate represents values of voxel intensity, the abscissa represents time (in seconds), zero corresponds to the onset of trials.

neutral in between. A similar pattern was observed for the superficial amygdala (SF), except that joy vs. neutral did not differ from each other in the right SF. In S1, fear evoked stronger BOLD signals than both neutral and joy (joy vs. neutral did not differ). In AC bilaterally, responses were stronger during the second half of each trial (and the same phenomenon was observed in area 45, the IFS, and the ventral pallidum/ventral striatum). By contrast, BOLD signals in the left SF were stronger during the first half. PPI results showed that both left and right AC showed stronger functional connectivity during joy

(compared to fear) with both the ipsilateral and the contralateral AC. During fear (compared to joy), both left and right AC showed stronger functional connectivity with the cuneus (areas 17 and 18), the median wall of the precuneus (areas 5 and 7), and almost the entire cingulate sulcus (CS). Moreover, both left and right AC showed stronger functional connectivity during fear with the anterior insula bilaterally, and the left (but not the right) AC showed stronger functional connectivity during fear with areas 3a & b of S1. The left SF showed stronger functional connectivity during joy (compared to fear) with right posterior Heschl's gyrus. During fear (compared to joy), the left SF showed stronger functional connectivity with cuneus (V1–V4), and area 7a of the superior parietal lobule (precuneus) bilaterally. The right SF showed stronger functional connectivity during joy (compared to fear) with the mediadorsal nucleus of the thalamus.

**Table 3**

Results of the interaction contrast of emotion (joy vs. fear)  $\times$  time (1st half vs. 2nd half of each trial), corrected for multiple comparisons ( $p < .05$ ). Abbreviations: AC: auditory cortex; FOp: frontal operculum; SF: superficial amygdala STG: superior temporal gyrus; l: left; r: right.

	MNI coordinate	Cluster size (mm <sup>3</sup> )	z-Value: max (mean)
l post. IFS	-51 18 31	1809	4.02 (3.39)
r post. IFS	45 21 34	1134	3.76 (3.17)
l pars triangularis (area 45) <sup>a</sup>	-51 30 10	-	3.76 (-)
r pars triangularis (area 45)	51 30 10	513	3.27 (3.26)
l ant insula/deep FOp	-34 34 4	621	4.22 (3.27)
l ant. STG	-60 -9 -2	1377	4.56 (3.45)
r ant. STG/planum polare	51 0 -8	2214	4.81 (3.40)
l planum temp. (AC)	-42 -30 13	2457	3.94 (3.35)
r planum temp. (AC)	45 -27 19	5589	4.50 (3.42)
l SF	-21 -6 -14	324	3.51 (3.19)
l pallidum	-15 0 7	1296	4.42 (3.47)
r pallidum	11 3 4	378	3.79 (3.31)

<sup>a</sup> The peak voxel in the l pars triangularis was part of the cluster with the maximum peak voxel in the l insula/deep frontal FOp.

#### Auditory cortex and emotional processing

Pronounced emotion-specific effects were observed in the auditory cortex: In the General Linear Model (GLM) contrast, BOLD responses in the entire supratemporal cortex (auditory core, belt, and parabelt) were stronger for joy than neutral stimuli, and stronger for neutral than fear stimuli. As will be argued in the following, these results indicate a prominent role of the auditory cortex in the emotional processing of auditory information. Importantly, there are five reasons as to why the activity differences between conditions cannot simply be due to acoustical factors: (1) the values of acoustical descriptors that significantly differed between conditions were included as covariates of no interest, and should therefore not have contributed to differences between conditions observed in the GLM contrasts. (2) However, even if

**Table 4**

Results of PPI analysis (corrected for multiple comparisons,  $p < .05$ ), separately for the seed voxels in: (a) left AC (Heschl's gyrus), (b) right AC (Heschl's gyrus), (c) left SF, and (d) right SF (the PPI analysis with S1 as seed region did not indicate any results). Positive  $z$ -values (outermost right column) indicate stronger functional connectivity during joy compared to fear, whereas negative  $z$ -values indicate stronger functional connectivity during fear compared to joy. Abbreviations: AC: auditory cortex; FOP: frontal operculum; HG: Heschl's gyrus; ITS: inferior temporal sulcus; MD: mediodorsal; MTG: middle temporal gyrus; PAC: primary auditory cortex; SF: superficial amygdala; SFS: superior frontal sulcus; SPL: superior parietal lobule; STG: superior temporal gyrus; l: left; r: right.

	MNI coordinate	Cluster size (mm <sup>3</sup> )	$z$ -Value: max (mean)
<b>(a) Left auditory cortex</b>			
l HG (PAC)	51 – 18 7	2349	4.26 (3.17)
r HG (PAC)	– 51 – 18 7	1431	3.77 (3.16)
r supramarginal gyrus	63 – 24 34	89,235	– 4.97 (– 3.22)
cuneus (area 17, 18)	9 – 66 4	19,764	– 3.99 (– 3.07)
l post. MTG/ITS	– 60 – 60 7	2646	– 4.19 (– 3.17)
l anterior insula	– 42 9 1	7263	– 4.52 (– 3.16)
l mid-insula <sup>a</sup>	– 41 – 5 14	–	– 3.5 (–)
l post. insula <sup>b</sup>	– 36 – 20 11	–	– 3.68 (–)
<b>(b) Right auditory cortex</b>			
l SFS	– 24 30 52	2997	– 4.07 (– 3.19)
l SFS	– 24 51 25	756	– 3.92 (– 3.04)
r SFS	27 42 37	3672	– 3.76 (– 3.05)
l ant. insula/deep FOP	– 45 12 1	648	– 4.06 (– 3.27)
l mid-insula	– 34 3 16	459	– 3.48 (– 3.05)
r ant. insula & putamen	27 12 7	2295	– 4.35 (– 3.17)
l planum temporale	– 60 – 24 7	2970	4.34 (3.22)
r planum temporale	60 – 18 4	1188	3.68 (3.13)
r supramarginal gyrus	63 – 42 37	3429	– 4.65 (– 3.17)
l post. MTG/ITS	– 60 – 66 7	3240	– 4.40 (– 3.24)
r post. MTG	54 – 57 13	5427	– 3.81 (– 3.06)
pre-genua cingulate	– 6 42 7	2268	– 3.64 (– 2.98)
cingulate sulcus	3 15 43	3888	– 4.38 (– 3.13)
post. cingulate sulcus	6 – 30 49	33,264	– 5.40 (– 3.18)
<b>(c) Left superficial amygdala</b>			
r planum temporale/post. HG	42 – 33 13	567	3.75 (3.05)
r SPL (area 7)	18 – 48 54	351	– 4.13 (– 2.94)
l SPL (area 7)	– 21 – 51 54	756	– 4.56 (– 3.21)
r sup. occipital gyrus (area 18)	26 – 95 20	7020	– 3.68 (– 2.87)
l middle occipital gyrus	– 30 – 78 19	11,043	– 4.24 (– 2.99)
l lingual gyrus (V4)	– 18 – 73 – 5	675	– 2.94 (– 2.69)
r lingual gyrus (V3 & V4)	18 – 81 – 8	648	– 2.85 (– 2.67)
<b>(d) Right superficial amygdala</b>			
MD thalamus	3 – 12 4	297	3.22 (2.93)

<sup>a</sup> The peak voxel in the l mid-insula was part of the cluster with the maximum peak voxel in the l ant. insula.

<sup>b</sup> The peak voxel in the l post.-insula was part of the cluster with the maximum peak voxel in the l ant. insula.

this procedure did not cancel out acoustical differences between conditions, joy and fear stimuli did not differ with regard to their intensity, mean FO frequency, variation of FO frequency, pitch centroid value, spectral complexity, and spectral flux. (3) FO salience and chord strength differed significantly between joy and fear stimuli, as well as between joy and neutral stimuli (FO salience was highest for neutral, intermediate for joy, and lowest for fear stimuli; chord strength was highest for joy stimuli, and did not differ between neutral and fear stimuli). Nevertheless, in the GLM, BOLD signal intensity in the auditory cortex was stronger in response to joy compared to neutral, and during neutral compared to fear stimuli; this pattern does not correlate with the pattern of FO salience (being strongest for neutral stimuli) or the pattern of chord strength (which did not differ between neutral and fear stimuli). (4) Key-strength showed differences between joy and fear, as well as between fear and neutral stimuli, but not between joy and neutral stimuli. Again, this pattern is not consistent with the pattern of BOLD responses observed in the auditory cortex. Although not well known, it is highly likely that extraction of the key of tonal music (including extraction of a tonal center) involves both posterior and anterior supratemporal cortex bilaterally (e.g., Koelsch, 2011; Liegeois-Chauvel et al., 1998; Patterson et al., 2002; Peretz and Zatorre, 2005). Therefore,

the interactions of the auditory cortex with limbic/paralimbic brain structures are likely to be due to emotional processes, rather than being merely due to cognitive processes related to the key-strength of sounds. (5) Although fear stimuli had a higher degree of sensory dissonance than joy stimuli, activity changes in the auditory cortex are unlikely to be due to this difference only, because neutral stimuli were even more consonant than joy stimuli. The pattern of BOLD signal intensity observed in the GLM contrast is, thus, not related to the degree of sensory dissonance of the stimuli.

Instead, the observed pattern of BOLD signal intensity in the AC corresponds to the emotion ratings for joy (and inversely for fear, respectively), indicating that activity of the auditory cortex is related to the emotional quality of auditory information: Compared to neutral, BOLD signals had a higher intensity during the joy condition, and a lower intensity during the fear condition. In other words, we observed an actual increase in BOLD activity during listening to joy stimuli and an actual decrease during listening to fear stimuli (compared to neutral stimuli). With regard to the pronounced regional activity in the auditory cortex during the joy-evoking music (as indicated by the GLMs), it is likely that this was in part due to a more detailed acoustical analysis of the joyful music, which was probably related to a voluntary shift of attention: participants had a preference for the joy stimuli (as indicated by the valence ratings), and therefore probably paid more voluntary attention to those stimuli, leading to a stronger auditory cortex activation (Jäncke et al., 1999). Similar findings have previously been reported for pleasant compared to unpleasant music (Koelsch et al., 2006; Mueller et al., 2011) or pleasant vs. unpleasant sounds from the International Affective Digitized Sound System (IADS, Plichta et al., 2011). However, it is unlikely that merely preference (and, correspondingly, voluntary shifts of attention) explains this effect, because the preference of participants was comparable between fear and neutral music (again, as indicated by the valence ratings), and yet BOLD signal intensity differed between fear and neutral.

The role of the auditory cortex in the emotional processing of auditory information is further highlighted by the PPI results involving auditory seed regions: These results revealed emotion-specific functional connectivity (a) between auditory cortical areas and cingulate, as well as insular cortex during joy stimuli, and (b) between auditory areas and parietal, as well as visual cortex (V1–V5) during fear stimuli. Both cingulate and insular cortex are involved in emotional processes, in particular with regard to autonomic regulation as well as the production of subjective feelings (Craig, 2009; Medford and Critchley, 2010). In addition, the cingulate cortex has been implicated in the coordination of autonomic activity, behavior, motor expression, as well as cognitive processes in response to emotionally salient stimuli (Koelsch et al., 2010b; Medford and Critchley, 2010).

With regard to the marked functional connectivity between auditory areas and parietal as well as visual cortex, anatomical studies indicate that core, belt and parabelt regions project to V1 and V2 of visual cortex, and that neurons in V2 project back into these auditory regions (reviewed in Smiley et al., 2007). The observed functional connectivity between these areas in the present study highlights the role of auditory-visual interactions, in particular during emotional states of fear. The functional significance of such interactions is probably increased visual alertness in the face of danger signaled by auditory information (probably including involuntary shifts of attention). Our results are the first to show that the auditory cortex is a central hub of an affective-attentional network that is more extensive than previously believed, involving functional connectivity of auditory association cortex with a diverse range of visual, attentional, and limbic/paralimbic structures. This finding also supports the notion that multisensory interactions in the cerebral cortex are not limited to established polysensory regions, but that "interactions with other sensory systems also take place in auditory cortex" (Smiley et al., 2007). Notably, this latter conclusion holds even if such multisensory interactions were due to acoustical features which were possibly not accounted for by the

computational feature extraction (and not necessarily related to emotional responses).

Many of the observed emotion-specific functional connections parallel anatomical connections previously described in monkeys (as described below). Our results provide information about the emotion-specific nature of such connections. With regard to functional connections to the insula, our results parallel connections between posterior AC and neighboring granular insula in macaque monkeys (Smiley et al., 2007), taken as a likely source of somatosensory input into the AC (Smiley et al., 2007). In addition, we observed functional connectivity not only with posterior, but also with mid- and anterior insula. This indicates clear functional connectivity between AC and the insula in humans, possibly reflecting sensory-limbic interactions that are more pronounced in humans than in monkeys. Such sensory-limbic interactions are also apparent in the extensive functional connectivity between AC and cingulate cortex. Previous studies with monkeys showed anatomical connections between (lateral) auditory belt and posterior cingulate cortex (Yukie, 1995). Our data suggest more extensive functional connections between auditory cortex and cingulate cortex in humans that also include anterior cingulate regions.

#### *Superficial amygdala and its role for joy and fear*

The superficial amygdala (SF) showed higher BOLD signal values bilaterally during joy compared to the fear stimuli. These findings corroborate previous reports of (right) SF activation in response to pleasant joyful music (compared to unpleasant music-like noise, Mueller et al., 2011). Due to its dense anatomical connections to the ventral striatum (from which it evolved phylogenetically, Nieuwenhuys et al., 2008), the superior amygdaloid complex has so far been implicated in positive emotion and hedonic processes (Nieuwenhuys et al., 2008), in line with our results. In addition, the superior amygdaloid complex has reciprocal connections to the orbitofrontal cortex (Bach et al., 2011) and plays a role for olfactory processes (Heimer and Van Hoesen, 2006; Price, 2003). Further functional connections include the caudate, cingulate cortex, insula, and hippocampus (Roy et al., 2009). Interestingly, a study by Goossens et al. (2009) suggested that the SF is particularly sensitive to social stimuli. Thus, in the present study, the joyful music possibly evoked activity within the SF due to the extraction of the social significance of the joyful music (but see also below). Such significance emerges from several social functions of music, including communication, coordination of movements, cooperation, and social cohesion (summarized in Koelsch, 2010). The fear stimuli, on the other hand, had no socially incentive value (being a signal of threat, and thus motivating withdrawal), probably resulting in decreased neuronal activity within the SF bilaterally (compared to joy and neutral stimuli). The fact that fear stimuli evoked significantly weaker responses in the right SF compared to a neutral control condition, and virtually no signal change in the left SF, suggests that the pattern of SF response to an auditory signal codes the emotional quality of that stimulus (i.e., whether the stimulus is an incentive social signal, or a signal of threat). Note that it is unlikely that SF simply codes valence (or arousal), for two reasons: *first*, to our knowledge, no previous study using stimuli that are perceived as rewarding, but do not have a social component (such as monetary rewards) reported SF activation, and *second*, valence as well as arousal ratings were used as regressors of no interest in the statistical modeling of the data, and are thus unlikely to contribute to the present fMRI results.

The PPI results reveal that functional connectivity between (left) SF and auditory regions was stronger during joy than during fear stimuli. Although previous studies have shown anatomical and functional connections between the basolateral (BL) amygdala and AC that are involved in fear conditioning (LeDoux, 2000), the significance of functional connectivity of the SF has remained elusive. As argued above, such connectivity is perhaps related to the social significance of stimuli, in contrast to the connectivity between BL and AC, which

appears to be important for the conditioning of (auditory) signals of danger. It has recently been proposed (Kumar et al., 2012) that amygdala activity affects AC activity as a function of the emotional valence of stimuli (and that AC provides limbic/paralimbic structures with information about the acoustic quality of sounds). Thus, the functional connectivity between (left) SF and AC observed in the present study is in part consistent with the results by Kumar et al. (2012), because the stronger AC activity during joy (compared to fear) might be related to amygdalar activity (note that the functional connectivity between SF and AC was stronger during joy than fear, and that joy also evoked stronger BOLD signals than fear in AC). The neural pathway that originates in SF and modulates AC activity remains to be specified; as will be discussed below, such a pathway probably involves thalamic nuclei, including the medio-dorsal thalamus. Notably, the study by Kumar et al. (2012) presented unpleasant stimuli only, thus our results suggest that amygdala activity is also related to AC activity in response to pleasant auditory stimuli.

In addition to joy, SF is also involved in fear responses, as indicated by the increased functional connectivity of the (left) SF with area 7 and with visual areas during fear (compared to joy), possibly related to the elicitation of increased visual alertness during fear-evoking auditory information. Finally, the right SF showed increased functional connectivity during joy with the medio-dorsal thalamus (MD). A diffusion-tensor-imaging study by Behrens et al. (2003) reported a fiber tract extending anteriorly and inferiorly along the medial wall of the thalamus, then turning laterally into the amygdala. A similar path has been documented for non-human primates, via the inferior thalamic peduncle (Aggleton and Mishkin, 1984). In the study by Behrens et al. (2003), this pathway was small, and the authors were thus not confident that their result was valid. However, our results suggest that this pathway from MD to the (superficial) amygdala exists, and that it plays a specific role for positive emotion. Perhaps this thalamic nucleus is part of the pathway by which AC activity is regulated as an effect of SF activity.

Contrary to our hypothesis, no activity changes were observed between conditions in the hippocampal formation. However, the activity changes observed in the SF spread into the hippocampal-amygdaloid transition area, and perhaps stronger signal changes in the hippocampal formation would have been obtained in a less noisy environment: Mueller et al. (2011) reported that significant signal changes in the hippocampal formation (evoked by pleasant joyful music contrasted to unpleasant music-like noise) were observed only with interleaved silent steady state scanning, or with sparse temporal scanning; no signal change was observed in the hippocampus during continuous scanning in that study.

#### *Primary somatosensory cortex (S1)*

Stronger BOLD signals were measured in right area 3b of S1 during fear than during joy (or neutral) in voxels that correspond to the cortical representation of the face in S1 (Blakemore et al., 2005; Moulton et al., 2009). Previous experiments have reported that the recognition of emotions from visually presented facial expressions requires right somatosensory-related cortices, including the face representation in S1 (Adolphs et al., 2000). That finding corroborated the notion that individuals recognize another individual's emotional state by internally generating somatosensory representations that simulate how the other individual would feel when displaying a certain emotional (facial) expression. Our data suggest that such somatosensory-driven simulations are also activated by auditory information with emotional valence, such as music (probably also affective prosody). This notion is consistent with data indicating facial mimicry in response to happiness or sadness expressed by music (Lundqvist et al., 2009). It is also possible that somatosensory activity reflects mapping of an evoked emotional state during the emergence of feelings with the aid of somatosensory representations (e.g., of proprioceptive information during

visually evoked emotions, Rudrauf et al., 2009). Again, our results suggest that such mapping can be activated by auditory information with emotional valence. The reason as to why, in our study, S1 representations were activated more strongly in response to fear than to joy remains to be specified.

## Conclusions

This study has two main conclusions: First, during music listening, the auditory cortex has emotion-specific functional interactions with a diverse range of visual, parietal, and limbic/paralimbic structures; this demonstrates that the auditory cortex is a central relay of an affective-attentional network that is more extensive than previously believed. This finding also implicates that the auditory cortex is involved in sensory-limbic and multisensory interactions that resemble those of established polysensory regions. Second, our results suggest that the superficial amygdala (SF) is sensitive for incentive social signals (including music), but at the same time also involved in fear responses: in concert with the auditory cortex, the SF appears to elicit increased visual alertness in the face of danger signaled by auditory information. Fear music may thus activate phylogenetically old mechanisms that engage the visual localization of potentially threatening objects. It is tempting to speculate that the corresponding increase of activity in visual areas during listening to fear-evoking music leads to more intense visual imagery (compared, e.g., to joyful music), particularly when listening to music with closed eyes (as in the present study). Such increased visual imagery during fear-evoking music might be an important factor contributing to the emotional experience, and the esthetic appeal, of fear-evoking music.

Supplementary data to this article can be found online at <http://dx.doi.org/10.1016/j.neuroimage.2013.05.008>.

## Acknowledgment

This research was funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) and the Cluster of Excellence “Languages of Emotion”.

## Conflict of interest

The authors declare that they do not have any conflict of interest.

## References

- Adolphs, R., Damasio, H., Tranel, D., Cooper, G., Damasio, A., 2000. A role for somatosensory cortices in the visual recognition of emotion as revealed by three-dimensional lesion mapping. *J. Neurosci.* 20, 2683–2690.
- Aggleton, J., Mishkin, M., 1984. Projections of the amygdala to the thalamus in the cynomolgus monkey. *J. Comp. Neurol.* 222, 56–68.
- Agrawal, D., Timm, L., Viola, F.C., Debener, S., Buechner, A., Dengler, R., Wittfoth, M., 2012. ERP evidence for the recognition of emotional prosody through simulated cochlear implant strategies. *BMC Neurosci.* 13, 113.
- Alluri, V., Toivainen, P., Jääskeläinen, I.P., Clerehugh, E., Sams, M., Brattico, E., 2012. Large-scale brain networks emerge from dynamic processing of musical timbre, key and rhythm. *NeuroImage* 59, 3677–3689.
- Amunts, K., Kedo, O., Kindler, M., Pieperhoff, P., Mohlberg, H., Shah, N., Habel, U., Schneider, F., Zilles, K., 2005. Cytoarchitectonic mapping of the human amygdala, hippocampal region and entorhinal cortex: intersubject variability and probability maps. *Anat. Embryol.* 210, 343–352.
- Bach, D., Behrens, T., Garrido, L., Weiskopf, N., Dolan, R., 2011. Deep and superficial amygdala nuclei projections revealed in vivo by probabilistic tractography. *J. Neurosci.* 31, 618–623.
- Baumgartner, T., Lutz, K., Schmidt, C., Jäncke, L., 2006. The emotional power of music: how music enhances the feeling of affective pictures. *Brain Res.* 1075, 151–164.
- Beck, A., Steer, R., Brown, G., 1993. Beck Depression Inventory. Psychological Corporation, San Antonio, TX.
- Behrens, T., Johansen-Berg, H., Woolrich, M., Smith, S., Wheeler-Kingshott, C., Boulby, P., Barker, G., Silbery, E., Sheehan, K., Ciccarelli, O., et al., 2003. Non-invasive mapping of connections between human thalamus and cortex using diffusion imaging. *Nat. Neurosci.* 6, 750–757.
- Blakemore, S., Bristow, D., Bird, G., Frith, C., Ward, J., 2005. Somatosensory activations during the observation of touch and a case of vision–touch synaesthesia. *Brain* 128, 1571–1583.
- Blood, A., Zatorre, R., 2001. Intensely pleasurable responses to music correlate with activity in brain regions implicated in reward and emotion. *Proc. Natl. Acad. Sci.* 98, 11818–11823.
- Boersma, P., 2002. Praat, a system for doing phonetics by computer. *Glott Int.* 5 (9/10), 341–345.
- Brattico, E., Alluri, V., Bogert, B., Jacobsen, T., Vartiainen, N., Nieminen, S., Tervaniemi, M., 2011. A functional MRI study of happy and sad emotions in music with and without lyrics. *Front. Psychol.* 2, 1–16.
- Brown, S., Martinez, M., Parsons, L., 2004. Passive music listening spontaneously engages limbic and paralimbic systems. *NeuroReport* 15, 2033–2037.
- Coutinho, E., Dibben, N., 2012. Psychoacoustic cues to emotion in speech prosody and music. *Cogn. Emot.* 1–27.
- Craig, A., 2009. How do you feel—now? The anterior insula and human awareness. *Nat. Rev. Neurosci.* 10, 59–70.
- Deichmann, R., Josephs, O., Hutton, C., Corfield, D., Turner, R., 2002. Compensation of susceptibility-induced bold sensitivity losses in echo-planar fMRI imaging. *NeuroImage* 15, 120–135.
- Deichmann, R., Gottfried, J., Hutton, C., Turner, R., 2003. Optimized EPI for fMRI studies of the orbitofrontal cortex. *NeuroImage* 19, 430–441.
- Eerola, T., Toivainen, P., 2004. Mir in matlab: the midi toolbox. Proceedings of the International Conference on Music Information Retrieval, pp. 22–27 (Citeseer).
- Eerola, T., Vuoskoski, J.K., 2011. A comparison of the discrete and dimensional models of emotion in music. *Psychol. Music.* 39, 18–49.
- Eickhoff, S., Stephan, K., Mohlberg, H., Grefkes, C., Fink, G., Amunts, K., Zilles, K., 2005. A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *NeuroImage* 25, 1325–1335.
- Ekman, P., 1999. Basic emotions. In: Dalglish, T., Power, M. (Eds.), *Handbook of Cognition and Emotion*. Wiley Online Library, pp. 45–60.
- Eldar, E., Ganor, O., Admon, R., Bleich, A., Hendler, T., 2007. Feeling the real world: limbic response to music depends on related content. *Cereb. Cortex* 17, 2828–2840.
- Friston, K., Buechel, C., Fink, G., Morris, J., Rolls, E., Dolan, R., 1997. Psychophysiological and modulatory interactions in neuroimaging. *NeuroImage* 6, 218–229.
- Friston, K., Ashburner, J., Kiebel, S., Nichols, T., Penny, W., 2007. *Statistical Parametric Mapping: the Analysis of Functional Brain Images*. Elsevier Academic Press, London.
- Fritz, T., Jentschke, S., Gosselin, N., Sammler, D., Peretz, I., Turner, R., Friederici, A.D., Koelsch, S., 2009. Universal recognition of three basic emotions in music. *Curr. Biol.* 19, 573–576.
- Gabrielson, A., Juslin, P., 2003. Emotional expression in music. In: Davidson, R., Scherer, K., Goldsmith, H. (Eds.), *Handbook of Affective Sciences*. Oxford University Press, New York, pp. 503–534.
- Geyer, S., Schleicher, A., Zilles, K., 1999. Areas 3a, 3b, and 1 of human primary somatosensory cortex: 1. microstructural organization and interindividual variability. *NeuroImage* 10, 63–83.
- Goossens, L., Kukolja, J., Onur, O., Fink, G., Maier, W., Griez, E., Schruers, K., Hurlemann, R., 2009. Selective processing of social stimuli in the superficial amygdala. *Hum. Brain Mapp.* 30, 3332–3338.
- Gosselin, N., Peretz, I., Noulhiane, M., Hasboun, D., Beckett, C., Baulac, M., Samson, S., 2005. Impaired recognition of scary music following unilateral temporal lobe excision. *Brain* 128, 628–640.
- Gosselin, N., Peretz, I., Johnsen, E., Adolphs, R., 2007. Amygdala damage impairs emotion recognition from music. *Neuropsychologia* 45, 236–244.
- Gosselin, N., Peretz, I., Hasboun, D., Baulac, M., Samson, S., 2011. Impaired recognition of musical emotions and facial expressions following anteromedial temporal lobe excision. *Cortex* 47, 1116–1125.
- Heimer, L., Van Hoesen, G., 2006. The limbic lobe and its output channels: implications for emotional functions and adaptive behavior. *Neurosci. Biobehav. Rev.* 30, 126–147.
- Huq, A., Bello, J.P., Rowe, R., 2010. Automated music emotion recognition: a systematic evaluation. *J. New Music Res.* 39, 227–244.
- Huron, D., 2006. *Sweet Anticipation: Music and the Psychology of Expectation*. The MIT Press.
- Janata, P., 2009. The neural architecture of music-evoked autobiographical memories. *Cereb. Cortex* 19, 2579–2594.
- Jäncke, L., Mirzazade, S., Shah, N.J., 1999. Attention modulates activity in the primary and the secondary auditory cortex: a functional magnetic resonance imaging study in human subjects. *Neurosci. Lett.* 266, 125–128.
- Johnstone, T., Ores Walsh, K., Greischar, L., Alexander, A., Fox, A., Davidson, R., Oakes, T., 2006. Motion correction and the use of motion covariates in multiple-subject fMRI analysis. *Hum. Brain Mapp.* 27, 779–788.
- Juslin, P., Laukka, P., 2003. Communication of emotions in vocal expression and music performance: different channels, same code? *Psychol. Bull.* 129, 770–814.
- Juslin, P., Västfjäll, D., 2008. Emotional responses to music: the need to consider underlying mechanisms. *Behav. Brain Sci.* 31, 559–575.
- Kim, Y.E., Schmidt, E.M., Migneco, R., Morton, B.G., Richardson, P., Scott, J., Speck, J.A., Turnbull, D., 2010. Music emotion recognition: a state of the art review. In *Proc. ISMIR*, pp. 255–266 (Citeseer).
- Koelsch, S., 2010. Towards a neural basis of music-evoked emotions. *Trends Cogn. Sci.* 14, 131–137.
- Koelsch, S., 2011. Towards a neural basis of music perception—a review and updated model. *Front. Psychol.* 2, 1–20.
- Koelsch, S., Fritz, T., Cramon, D.Y., Müller, K., Friederici, A.D., 2006. Investigating emotion with music: an fMRI study. *Hum. Brain Mapp.* 27, 239–250.
- Koelsch, S., Offermanns, K., Franzke, P., 2010. Music in the treatment of affective disorders: an exploratory investigation of a new method for music–therapeutic research. *Music. Percept.* 27, 307–316.

- Koelsch, S., Siebel, W.A., Fritz, T., 2010. Functional neuroimaging of emotion with music. In: Juslin, P., Sloboda, J. (Eds.), *Handbook of Music and Emotion: Theory, Research, Applications*. Oxford University Press Oxford, Oxford, pp. 313–346.
- Koelsch, S., Fuernmetz, J., Sack, U., Bauer, K., Hohenadel, M., Wiegel, M., Kaisers, U., Heinke, W., 2011. Effects of music listening on cortisol levels and propofol consumption during spinal anesthesia. *Front. Psychol.* 2, 1–9.
- Krumhansl, C., 1990. *Cognitive Foundations of Musical Pitch*. Oxford University Press, USA.
- Kumar, S., von Kriegstein, K., Friston, K., Griffiths, T.D., 2012. Features versus feelings: dissociable representations of the acoustic features and valence of aversive sounds. *J. Neurosci.* 32, 14184–14192.
- Laurier, C., 2011. *Automatic Classification of Musical Mood by Content-Based Analysis*. Ph.D. thesis Universitat Pompeu Fabra, Barcelona.
- LeDoux, J., 2000. Emotion circuits in the brain. *Ann. Rev. Neurosci.* 23, 155–184.
- Lerner, Y., Papo, D., Zhdanov, A., Belozersky, L., Hendler, T., 2009. Eyes wide shut: amygdala mediates eyes-closed effect on emotional experience with music. *PLoS One* 4, e6230.
- Liegeois-Chauvel, C., Peretz, I., Babiak, M., Laguitton, V., Chauvel, P., 1998. Contribution of different cortical areas in the temporal lobes to music processing. *Brain* 121, 1853–1867.
- Lohmann, G., Müller, K., Bosch, V., Mentzel, H., Hessler, S., Chen, L., von Cramon, D.Y., 2001. *Lipsia* – a new software system for the evaluation of functional magnet resonance images of the human brain. *Comput. Med. Imaging Graph.* 25, 449–457 (See also at <http://www.cns.mpg.de/lipsia>).
- Lohmann, G., Neumann, J., Müller, K., Lepsius, J., Turner, R., 2008. The multiple comparison problem in fMRI – a new method based on anatomical priors. *Workshop on Analysis of Functional Medical Images*, New York University, pp. 179–187.
- Lundqvist, L., Carlsson, F., Hilmersson, P., Juslin, P., 2009. Emotional responses to music: experience, expression, and physiology. *Psychol. Music.* 37, 61–90.
- Mattheson, J., 1739/1999. *Der vollkommene Capellmeister*. Bärenreiter, London.
- Medford, N., Critchley, H., 2010. Conjoint activity of anterior insular and anterior cingulate cortex: awareness and response. *Brain Struct. Funct.* 214, 535–549.
- Menon, V., Levitin, D., 2005. The rewards of music listening: response and physiological connectivity of the mesolimbic system. *NeuroImage* 28, 175–184.
- Menon, V., Levitin, D., Smith, B., Lemke, A., Krasnow, B., Glazer, D., Glover, G., McAdams, S., 2002. Neural correlates of timbre change in harmonic sounds. *NeuroImage* 17, 1742–1754.
- Mitterschiffthaler, M.T., Fu, C.H., Dalton, J.A., Andrew, C.M., Williams, S.C., 2007. A functional MRI study of happy and sad affective states evoked by classical music. *Hum. Brain Mapp.* 28, 1150–1162.
- Morosan, P., Rademacher, J., Schleicher, A., Amunts, K., Schormann, T., Zilles, K., 2001. Human primary auditory cortex: cytoarchitectonic subdivisions and mapping into a spatial reference system. *NeuroImage* 13, 684–701.
- Moulton, E., Pendse, G., Morris, S., Aiello-Lammens, M., Becerra, L., Borsook, D., 2009. Segmentally arranged somatotopy within the face representation of human primary somatosensory cortex. *Hum. Brain Mapp.* 30, 757–765.
- Mueller, K., Mildner, T., Fritz, T., Lepsius, J., Schwarzbauer, C., Schroeter, M., Möller, H., 2011. Investigating brain response to music: a comparison of different fMRI acquisition schemes. *NeuroImage* 54, 337–343.
- Mutschler, I., Wieckhorst, B., Speck, O., Schulze-Bonhage, A., Hennig, J., Seifritz, E., Ball, T., 2010. Time scales of auditory habituation in the amygdala and cerebral cortex. *Cereb. Cortex* 20, 2531–2539.
- Nieuwenhuys, R., Voogd, J., Huijzen, C.V., 2008. *The Human Central Nervous System*. Springer, Berlin.
- Patterson, R., Uppenkamp, S., Johnsrude, I., Griffiths, T., 2002. The processing of temporal pitch and melody information in auditory cortex. *Neuron* 36, 767–776.
- Peretz, I., 2010. Towards a neurobiology of musical emotions. In: Juslin, P., Sloboda, J. (Eds.), *Handbook of Music and Emotion: Theory, Research, Applications*. Oxford University Press, Oxford, pp. 99–126.
- Peretz, I., Zatorre, R., 2005. Brain organization for music processing. *Ann. Rev. Psychol.* 56, 89–114.
- Petrides, M., Pandya, D., 1988. Association fiber pathways to the frontal cortex from the superior temporal region in the rhesus monkey. *J. Comp. Neurol.* 273, 52–66.
- Phan, K., Wager, T., Taylor, S., Liberzon, I., 2002. Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in PET and fMRI. *NeuroImage* 16, 331–348.
- Plichta, M., Gerdes, A., Alpers, G., Harnisch, W., Brill, S., Wieser, M., Fallgatter, A., 2011. Auditory cortex activation is modulated by emotion: a functional near-infrared spectroscopy (fNIRS) study. *NeuroImage* 55, 1200–1207.
- Plomp, R., Levelt, W., 1965. Tonal consonance and critical bandwidth. *J. Acoust. Soc. Am.* 38, 548–560.
- Price, J., 2003. Comparative aspects of amygdala connectivity. *Ann. N. Y. Acad. Sci.* 985, 50–58.
- Roy, A., Shehzad, Z., Margulies, D., Kelly, A., Uddin, L., Gotimer, K., Biswal, B., Castellanos, F., Milham, M., 2009. Functional connectivity of the human amygdala using resting state fMRI. *NeuroImage* 45, 614–626.
- Rudrauf, D., Lachaux, J., Damasio, A., Baillet, S., Hugueville, L., Martinerie, J., Damasio, H., Renault, B., 2009. Enter feelings: somatosensory responses following early stages of visual induction of emotion. *Int. J. Psychophysiol.* 72, 13–23.
- Salimpoor, V., Benovoy, M., Larcher, K., Dagher, A., Zatorre, R., 2011. Anatomically distinct dopamine release during anticipation and experience of peak emotion to music. *Nat. Neurosci.* 14, 257–262.
- Scheperjans, F., Eickhoff, S., Hömke, L., Mohlberg, H., Hermann, K., Amunts, K., Zilles, K., 2008. Probabilistic maps, morphometry, and variability of cytoarchitectonic areas in the human superior parietal cortex. *Cereb. Cortex* 18, 2141–2157.
- Smiley, J., Hackett, T., Ulbert, I., Karmas, G., Lakatos, P., Javitt, D., Schroeder, C., 2007. Multisensory convergence in auditory cortex, i. cortical connections of the caudal superior temporal plane in macaque monkeys. *J. Comp. Neurol.* 502, 894–923.
- Stein, J., Wiedholz, L., Bassett, D., Weinberger, D., Zink, C., Mattay, V., Meyer-Lindenberg, A., 2007. A validated network of effective amygdala connectivity. *NeuroImage* 36, 736–745.
- Trost, W., Ethofer, T., Zentner, M., Vuilleumier, P., 2012. Mapping aesthetic musical emotions in the brain. *Cereb. Cortex* 22 (12), 2769–2783.
- Vassilakis, P.N., Kendall, R.A., 2010. Psychoacoustic and cognitive aspects of auditory roughness: definitions, models, and applications. *Proceedings of Human Vision and Electronic Imaging XV*, 7527, pp. 1–7.
- Vieillard, S., Peretz, I., Gosselin, N., Khalfa, S., Gagnon, L., Bouchard, B., 2008. Happy, sad, scary and peaceful musical excerpts for research on emotions. *Cogn. Emot.* 22, 720–752.
- Weiskopf, N., Hutton, C., Josephs, O., Turner, R., Deichmann, R., 2007. Optimized EPI for fMRI studies of the orbitofrontal cortex: compensation of susceptibility-induced gradients in the readout direction. *Magn. Reson. Mater. Phys., Biol. Med.* 20, 39–49.
- Williams, L., Das, P., Liddell, B., Kemp, A., Rennie, C., Gordon, E., 2006. Mode of functional connectivity in amygdala pathways dissociates level of awareness for signals of fear. *J. Neurosci.* 26, 9264–9271.
- Yukie, M., 1995. Neural connections of auditory association cortex with the posterior cingulate cortex in the monkey. *Neurosci. Res.* 22, 179–187.



## 4. THE AGE OF CONTEXT-AWARENESS

*My cow is not pretty, but it's pretty to me*

David Lynch

*Well, shall we think or listen?  
Is there a sound addressed not wholly to the ear?  
We half close our eyes.  
We do not hear it through our eyes.  
It is not a flute note either,  
it is the relation of a flute note to a drum.*

William Carlos Williams, *The Orchestra*

### 4.1. Introduction

“An essential part of human psychology is the ability to identify music, text, images or other information based on associations provided by contextual information of different media” (Brochu et al., 2003). Because of that, studies abound on the different usages of music and their contexts of listening (Schedl et al., 2012; Stober, 2011). But context was, indeed, one of the missing elements in many of the early approaches to MIR and a partial cause of the semantic gap (see the previous chapter) (Wiggins, 2009; Schedl & Knees, 2009). With the availability of portable personal communication devices, which include sensing capabilities, the possibility of detecting and characterizing contextual data was made feasible and, hence, a new series of research and applied questions was posed (Kaminskas & Ricci, 2009; Schedl, 2013), marking the beginning of the age of context-awareness.

It is difficult to find a single notion or definition of context (but see (Dey & Abowd, 2000; Dey, 2001) because different types of contexts have been attracted the attention of researchers:

- Physiological parameters, which can be altered as a consequence of music listening (Bernardi et al., 2009) and hence can be used to track and modulate pleasantness or discomfort (Thaut & Davis, 1993; Kallinen & Ravaja, 2004; Nagel, 2007; Bernatzky et al., 2011; Knox et al., 2011), or to track musical sections, motives or special musical events (Cabredo et al., 2011; Oliver & Kreger-Stickles, 2006; Masahiro et al., 2008). Alternatively, these parameters can act as triggers for certain musical selections and for musical behaviour, so they provide context but also they are affected by it (Hu et al., 2006).
- The listener itself, as a carrier of a listening history, music preferences or personality, and subject to a short-term changing mindset and mood (Baur et al, 2010; Baur, 2011; Schedl et al., 2013). Mindset and mood may act like high-level surrogates of the physiological parameters, with the advantage that they are usually reported verbally, not requiring specific tracking devices (but they can be automatically inferred, up to a certain point, from measures taken by them)

(Schubert, 2002). The connection between personality and music preferences have been successfully explored and characterized (Rentfrow & Gosling, 2003; Rentfrow et al., 2011) and some applications start to take advantage of it (Fernández-Tobias et al., 2016; Ferwerda et al., 2016). An important but still open issue is how the emotional content of music can modify the emotional states or mood of the listener. Predictive models of that are still missing or are rudimentary (Han et al., 2010; Gilda, 2017).

- Activities in which listeners play or listen to music. The most usual is working-out, but studying, dancing, getting asleep, driving or making love are also common in listener's reports (Hu et al., 2006; Lee & Downie, 2004). Portable music devices addressed to them have been created (one of the first appliances incorporating automatic music audio analysis was Yamaha's BodyBeat<sup>30</sup>, targeted to people working out with music tempos matching the rhythm of the jogger). Behavioural measures like mouse-clicks, button pressings, playlist skips, stepping or writing velocity, etc., are also used to infer part of that context (Pampalk, Pohle and Widmer, 2005; Stober, 2011; Hu & Ogihara, 2011).
- Physical environment context, including temporal, geographical and weather conditions: listening to music can be geo-tagged in order to provide concert-going suggestions, and time of listening can be used as a predictor of future listening (Lee & Lee, 2006; Baltrunas & Amatriain, 2009; Herrera et al., 2010). Environmental noise, illumination or weather conditions, being factors that affect human behaviour, can also be sensed and used to modulate musical recommendations, playlists or playing parameters (Park et al., 2006).
- Peers: social context is a powerful modulator of preferences and opinions about music and artists (Hargreaves & North, 1999; Uitdenbogerd & Van Steelant, 2002; Fields et al., 2008; Barrington et al., 2009; Liu & Reimer, 2008). For this reason, computing similarities and relatedness, or even basing recommendation systems on peer-based data, such as tags from social websites, have been proposed (Levy & Sandler, 2007; Lamere, 2008; Shavitt & Weinsberg, 2009).
- Culture: we now know that even before birth the music culture surrounding us exerts a crucial role by means of implicit learning of musical regularities (Tillmann & Bharucha, 2000) and hence our familiarity or preference for certain musical features is influenced by those that are predominant there. Such cultural connections, stereotypes or assumptions can be inferred by web-mining, playlist-mining or social-mining and be used to filter retrieval, guide searches or build similarities between tracks, artists, contexts of use, etc. (Whitman & Lawrence, 2002; Baumann & Hummel, 2005; Knees et al., 2008; Schedl, 2008; Schedl et al., 2008; Levy & Sandler, 2008).
- The context music itself sets. This type of context is the only one being purely musical as it comprises the music happening before (and sometimes after) a specific moment in time. Purely musical context has a crucial role for adaptive

---

<sup>30</sup> <http://www.yamaha.com/bodibeat>

signal processing<sup>31</sup> and feature extraction algorithms (Goto, 2001; Müller et al., 2009; Peeters, 2006; Stober, 2011). Intrinsic musical context is very relevant for playlist continuation (Fields, 2011; Bonnin & Jannach, 2014; Vall et al., 2017), a fruitful and somehow new topic that motivated a challenge in the context of the RecSys conference<sup>32</sup>. Most of the approaches rely on the most recent listening only, hence overlooking the benefits that long-term listening patterns might exert (Kamehkosh et al., 2016).

- Music-synchronous or visual-linked media. In addition to lyrics analysis (Hyung et al., 2014), video accompanying music (Schindler & Rauber, 2015), record sleeves (Libeks & Turnbull, 2011) and artists' photos (Libeks & Turnbull, 2010), provide additional data streams to compute similarities, describe preferences or suggest music. "Multi-modality" becomes then a solid research topic in MIR (Mayer & Rauber, 2010; Schedl & Knees, 2011, Oramas et al., 2017).

If there was one type of application that probably reaches a maturity status in this age, it could be that of recommender systems and therefore we can witness the flourishing of a myriad of music recommenders that were taking advantage of contextual information and hence, improving the figures of merit over pure audio-based ones (Baltrunas et al., 2011; Hu & Ogihara, 2011; Wang et al., 2012; Kaminskas et al., 2013; Cheng & Shen, 2014; Schedl et al., 2014).

Understanding music-related context has called for knowledge from disciplines as diverse as biology, neuroscience, psychology, linguistics, ergonomics or economics. Surprisingly, the applied intentions of many MIR researchers did not find in those disciplines the answers for some of their questions and needs, and this has fostered the inherent interdisciplinarity in the field.

## 4.2. Papers included in this chapter

Bogdanov, D., Haro, M., Fuhrmann, F., Xambó, A., Gómez, E. & **Herrera, P.** (2013) Semantic content-based music recommendation and visualization based on user preference examples. *Information Processing and Management*, 49(1), 13-33. (h-index: 84; Journal IF 2017: 3.444, Q1 in Information Processing journals; 71 citations)

**Herrera, P.**, Resa Z., & Sordo M. (2010). *Rocking around the clock eight days a week: an exploration of temporal patterns of music listening*. 1st Workshop on Music Recommendation and Discovery (WOMRAD), ACM RecSys, 2010, Barcelona, Spain. (27 citations, WIRED magazine short note, idea quickly adopted by last.fm<sup>33</sup>)

---

<sup>31</sup> See also Celemony System's Audio Random Access technology, tying traditional signal processing with audio content analysis to improve the former in "real-time".

[https://en.wikipedia.org/wiki/Audio\\_Random\\_Access](https://en.wikipedia.org/wiki/Audio_Random_Access)

<sup>32</sup> <http://2018.recsyschallenge.com>

<sup>33</sup> <http://blog.last.fm/2010/09/06/now-in-the-playground-listening-clocks>

### 4.3. Contributions in the selected papers

In "Semantic content-based music recommendation and visualization based on user preference examples" we built individualized user models to avoid a cold start with a music recommender. Preference elicitation is a challenging fundamental problem when designing recommender systems, so we based our recommender on an explicit set of music tracks provided by the involved users as "representative" of their musical tastes and interests. We inferred from them a user model based on semantic descriptors (i.e., if they preferred danceable or non-danceable, vocal or instrumental, tonal or atonal, fast or slow music, etc.). Then we showed the utility of such profile for, on one side, automatically creating a cartoon humanoid or "musical avatar", as a device to visualize and socially communicate music preferences (different body and clothing parts changed accordingly to preferred aspects). Our subjects judged their respective avatars as good and compact ways to capture their preferences. On the other hand, the user profile was used to recommend music by means of three different strategies, two of them based on a semantic music similarity measure (computed either using just one of the preferred items or using all of them), and one strategy based on a Gaussian Mixture Model computed using the whole preferred set. We used qualitative dimensions such as familiarity with the recommended items, liking of them, and intentions (to further listening to them), to evaluate the users' satisfaction with the recommendation. Interestingly, in addition to hits (unfamiliar tracks that were liked and intended to be further listened to) and fails (disliked tracks with no future listening prospects) we introduced the category "trust" to consider tracks that, although familiar (hence not "discoveries") were liked and intended for further listening. Trusts are indicators that the system "understands" the tastes of the user, and a moderate number of them, scattered among many hits, seemed to be a desirable feature (in fact Last.fm supplied 25% of them whereas our tested methods supplied less than 4% trusted items). Overall our semantic-descriptor-based recommendations were better than those based on low-level features only, and better than those generated using just genre labels. Even though our fails were higher than the hits, it is remarkable for an audio content-based recommender to get hits just 7 percent units below than a full-fledged commercial system as Last.fm. Qualitatively, the user profile computed from a preference set assures a noise-free representation of the user's preference with a maximum possible coverage. Moreover, the chosen preference elicitation technique -namely inferring the dimensions of the user's preferences in a fully audio content-based manner- affords the system to overcome the so-called cold-start problem, which audio content-unaware systems must face. It also guarantees recommendations of non-popular items, which may be preferred by specialized or long-tail seeking listeners. Finally, having semantic descriptions for both the user and the recommended items allows to automatically generate justifications for the recommended music (e.g., "This track was recommended because you like jazz music with acoustic instrumentation and relaxed mood".), which is a highly desirable feature for a recommendation system (Tintarev & Masthoff, 2007).

In "Rocking around the clock eight days a week: an exploration of temporal patterns of music listening" we presented one of the earliest studies (if not the first one) on how music listening patterns can be influenced by contextual factors such as the day of the week or the time where listening happens. Chronobiology has demonstrated in the last 50 years that there are many biological processes (and hence, probably behavioural patterns) that are driven or, at least, modulated by inner and outer clocks. Music listening could be just another example, although the nature and method of our research cannot demonstrate

any causal relation. Here we addressed the hypothesis that, for some listeners, certain moments of the day or certain days of the week could yield a clear preference for some artists or genres. With the help of circular statistics (which was also an innovation in the MIR literature), we analysed playcounts from Last.fm (thanks to the dataset made publicly available by Celma

<sup>34</sup>) and we detected the existence of that kind of patterns. Once temporal preference was modelled for each listener, we tested the robustness of it using the listener's playcount from a posterior temporal period. We showed that, for some users, artists and genres, temporal patterns of listening could be used to predict music listening selections with above-chance accuracy. This finding could be exploited in music recommendation and playlist generation to provide user-specific music suggestions at the “right” moment. Research on the topic has been updated with expanded focus, including country and other demographic information (Schedl, 2017), focus on seasonal effects (Pettijohn et al., 2010; Krause & North, 2018), or weather conditions (Karmaker et al., 2018).

---

<sup>34</sup> <http://www.dtic.upf.edu/~ocelma/MusicRecommendationDataset/lastfm-1K.html>

Bogdanov, D., Haro, M., Fuhrmann, F., Xambó, A., Gómez, E. & **Herrera, P.** (2013) "Semantic content-based music recommendation and visualization based on user preference examples". *Information Processing and Management*, 49(1), 13-33.

DOI:<https://doi.org/10.1016/j.ipm.2012.06.004>

ISSN: 0306-4573



Contents lists available at SciVerse ScienceDirect

## Information Processing and Management

journal homepage: [www.elsevier.com/locate/infoproman](http://www.elsevier.com/locate/infoproman)

## Semantic audio content-based music recommendation and visualization based on user preference examples

Dmitry Bogdanov<sup>a,\*</sup>, Martín Haro<sup>a</sup>, Ferdinand Fuhrmann<sup>a</sup>, Anna Xambó<sup>b</sup>, Emilia Gómez<sup>a</sup>, Perfecto Herrera<sup>a</sup>

<sup>a</sup> Music Technology Group, Universitat Pompeu Fabra, Roc Boronat 138, 08018 Barcelona, Spain

<sup>b</sup> Music Computing Lab, The Open University, Walton Hall, MK7 6AA Milton Keynes, UK

### ARTICLE INFO

#### Article history:

Received 26 September 2011

Received in revised form 15 February 2012

Accepted 17 June 2012

Available online 25 July 2012

#### Keywords:

Music information retrieval

Information systems

User modeling

Recommender system

Preference visualization

Evaluation

### ABSTRACT

Preference elicitation is a challenging fundamental problem when designing recommender systems. In the present work we propose a content-based technique to automatically generate a semantic representation of the user's musical preferences directly from audio. Starting from an explicit set of music tracks provided by the user as evidence of his/her preferences, we infer high-level semantic descriptors for each track obtaining a user model. To prove the benefits of our proposal, we present two applications of our technique. In the first one, we consider three approaches to music recommendation, two of them based on a semantic music similarity measure, and one based on a semantic probabilistic model. In the second application, we address the visualization of the user's musical preferences by creating a humanoid cartoon-like character – the *Musical Avatar* – automatically inferred from the semantic representation. We conducted a preliminary evaluation of the proposed technique in the context of these applications with 12 subjects. The results are promising: the recommendations were positively evaluated and close to those coming from state-of-the-art metadata-based systems, and the subjects judged the generated visualizations to capture their core preferences. Finally, we highlight the advantages of the proposed semantic user model for enhancing the user interfaces of information filtering systems.

© 2012 Elsevier Ltd. All rights reserved.

### 1. Introduction

Over the past decade, we have witnessed a rapid growth of digital technologies, the Internet, and the multimedia industry. Consequently, the overload of generated information has created the current need for effective information filtering systems. Such information systems include tools for browsing and indexing large data catalogs as well as recommendation algorithms to discover unknown but relevant items therein. Their development and related research are usually carried out in the field of information retrieval.

In particular, recommender systems built upon user profiles are currently in the spotlight of the information retrieval community. Since preferences are highly subjective, personalization seems to be a key aspect for optimal recommendation. Ideally, such systems should be able to grasp user preferences and provide, on this basis, the content which is relevant to the user's needs.

Preference elicitation can therefore be regarded as a fundamental part of recommender systems and information filtering systems in general. Several approaches have been proposed in the literature to tackle this problem. In particular, Hanani,

\* Corresponding author. Tel.: +34 935422000/935422164; fax: +34 935422517.

E-mail address: [dmitry.bogdanov@upf.edu](mailto:dmitry.bogdanov@upf.edu) (D. Bogdanov).

Shapira, and Shoval (2001) identified two main strategies – explicit and implicit user preference inference. The former relies on user surveys in order to obtain qualitative statements and ratings about particular items or more general semantic properties of the data. In contrast, the latter relies on the information inferred implicitly from user behavior and, in particular, consumption statistics. In the present work, we focus on music recommender systems and consider explicit strategies to infer musical preferences of a user directly from the music audio data.

When considering digital music libraries, current major Internet stores contain millions of tracks. This situation complicates the user's search, retrieval, and discovery of relevant music. At present, the majority of industrial systems provide means for manual search (Nanopoulos, Rafailidis, Ruxanda, & Manolopoulos, 2009). This type of search is based on metadata<sup>1</sup> information about artist names, album or track titles, and additional semantic<sup>2</sup> properties which are mostly limited to genres. Music collections are then queried by tags or textual input using this information.

Moreover, current systems also provide basic means for music recommendation and personalization, which are not related to the audio content, i.e., using metadata. Such systems obtain a user's profile by monitoring music consumption and listening statistics, user ratings, or other types of behavioral information, decoupled from the actual music data (Baltrunas & Amatriain, 2009; Celma, 2008; Firan, Nejdil, & Paiu, 2007; Jawaheer, Szomszor, & Kostkova, 2010; Levy & Bosteels, 2010; Shardanand & Maes, 1995). In particular, a user can be simply represented as a vector of ratings or playback counts for different artists, albums, and tracks. Having a database of such user profiles, this allows the use of collaborative filtering to search for similar users or music items (Sarwar, Karypis, Konstan, & Reidl, 2001). Alternatively, semantic tag-based profiles can be built to be matched with music items directly. Firan et al. (2007) proposes to create such a semantic profile using implicit information about a user's listening behavior. To this end, they use the user's listening statistics (artist or track playback counts) and the editorial metadata extracted from the files in the user's personal music collection (artist names and track titles). The tags are obtained for artists, albums, and particular tracks from music services which provide means for social tagging, such as *Last.fm*.<sup>3</sup> Tags can also be retrieved from information found on the Web (Celma, 2008; Celma & Serra, 2008; Schedl, Widmer, Knees, & Pohle, 2011) in the form of reviews, biographies, blog posts, music related RSS feeds, etc.

These approaches, notably using collaborative filtering for music recommendation, are found to be effective when considering popular music items. However, it has been shown that they fail in the long tail, i.e., for unpopular items, due to the lack of available user ratings, social tags, and other types of metadata (Celma, 2008). On the other hand, there is evidence (Barrington, Oda, & Lanckriet, 2009; Celma & Herrera, 2008) that content-based<sup>4</sup> information extracted from the audio can help to overcome this problem.

Existing research in the area of audio content-based music recommendation usually focuses on the related task of measuring music similarity. The Music Information Research (MIR) community has achieved relative success in this task (Casey et al., 2008; Downie, Ehmann, Bay, & Jones, 2010), striving to facilitate both manual search and automatization of music recommendation. In these approaches, music tracks are represented in a given feature space, built upon timbral, temporal, tonal, and/or higher-level semantic dimensions, all extracted from audio content (Barrington, Turnbull, Torres, & Lanckriet, 2007; Bogdanov, Serrà, Wack, & Herrera, 2009; Pampalk, 2006; Pohle, Schnitzer, Schedl, Knees, & Widmer, 2009; West & Lamere, 2007). Such a representation enables the definition of similarity measures (or distances<sup>5</sup>) between tracks, which can be used to search music collections using queries-by-example. Such distance-based approaches are designed and evaluated, in most cases, for the query-by-one-example use-case. Since retrieval based on a single example is just a particular case of using a recommender system, these approaches may not be directly suitable for music recommendation purposes in general. As the users only provide one query, no knowledge about their musical preferences is required. Querying by example implies an active interaction by the user to explicitly define the “direction of search”. As a result, such approaches are not suitable when a user does not know her/his exact needs and prefers receiving recommendations from an available music collection without defining an example (seed) item.

In addition to these non-personalized measures, there has only been sparse work on personalized music similarity measures from audio content data (Lu & Tseng, 2009; Sotiropoulos, Lampropoulos, & Tsihrintzis, 2007; Vignoli & Pauws, 2005). These studies introduce metrics, which are adapted according to a user's perception of similarity to measure distances between tracks in a given collection. Nevertheless, these studies are also focused on the query-by-one-example scenario, and, in their majority, do not take musical preferences into account.

Alternatively, there exist few research studies on user preference modeling for music recommendation which include studies of audio content-based (Grimaldi & Cunningham, 2004; Hoashi, Matsumoto, & Inoue, 2003; Logan, 2004; Mandel & Ellis, 2005) and hybrid approaches (Li, Myaeng, Guan, & Kim, 2005; Su, Yeh, & Tseng, 2010; Yoshii, Goto, Komatani, Ogata, & Okuno, 2006). These studies present several shortcomings. Firstly, they operate solely on rough timbral, and sometimes temporal and tonal information. This information is low-level as it does not incorporate higher-level semantics in the description of music. In the case of music similarity, it has been shown that distance measures which operate on semantic descriptors, inferred from low-level features, outperform low-level derived similarities (Barrington et al., 2007; Bogdanov

<sup>1</sup> We pragmatically use the term “metadata” to refer to any information not extracted from the audio signal itself.

<sup>2</sup> We use the term “semantic” to refer to the concepts that music listeners use to describe items within music collections, such as genres, moods, musical culture, and instrumentation.

<sup>3</sup> <http://last.fm>.

<sup>4</sup> We use the terms “audio content-based” or “audio content” to refer to any information extracted from the raw audio signal.

<sup>5</sup> For the sake of simplicity we refer to any (dis) similarity estimation with the term “distance”.



et al., 2009; West & Lamere, 2007). Recent research suggests that exploiting a semantic domain can be a relevant step to overcome the so-called semantic gap (Aucouturier, 2009; Celma, Herrera, & Serra, 2006), which arises from the weak linking between human concepts related to musical aspects and the low-level feature data extracted from the audio signal. Furthermore, the metadata components of the majority of hybrid approaches solely use information about user ratings, exploiting it in a collaborative filtering manner. This allows to measure relations between different music tracks or between different users, but does not provide insights into the underlying relations between the music and the user himself, i.e., the nature of musical preferences. Moreover, a large amount of users and ratings is usually required for reasonable performance, as such systems are prone to the so-called “cold-start problem” (Maltz & Ehrlich, 1995), i.e., the inability to provide good recommendations at the initial stages of the system.

This all indicates a lack of research on both metadata-based and audio content-based strategies for an effective elicitation of musical preferences, including comprehensive evaluations on large music collections and real listeners. Most existing approaches exploit user ratings as the only source of explicit information. The evaluation of such approaches is often done objectively without the participation of real listeners. Ground truth datasets of user ratings are used instead. However, these ratings can be considered as indirect and even noisy preference statements (Amatriain, Pujol, & Oliver, 2009). They do not necessarily represent real user preferences, as they are biased by the precision of a rating scale, decisions on the design of the recommender interface, etc. (Cosley, Lam, Albert, Konstan, & Riedl, 2003). In turn, implicit listening behavior statistics based on track counts might not represent real preferences in particular since it ignores the difference between track durations or users' activities when listening the music (Jawaheer et al., 2010). Furthermore, these information sources do not guarantee a complete coverage of all kinds of preferred items. Alternative explicit approaches are generally limited to surveying for the names of favorite artists, albums, or preferred genres.

In the present work, we focus on audio content-based user modeling suitable for music recommendation. In contrast to most existing approaches, we propose a novel technique which is based on the automatic inference of a high-level semantic description<sup>6</sup> of the music audio content, covering different musical facets, such as genre, musical culture, moods, instruments, rhythm, and tempo. These semantic descriptors are computed from an explicit set of music tracks defined by a given user as evidence of her/his musical preferences. To the best of our knowledge this approach for user modeling for music recommendation has never been evaluated before. In particular, our technique relies on two hypotheses. First, we suppose that asking for explicit preference examples is an effective way to infer real user preferences. Second, we assume that high-level semantic description outperforms common low-level feature information in the task of music recommendation. The latter hypothesis is based on similar evidence in the case of music similarity estimation (Bogdanov et al., 2009).

In particular, our focus lies on music discovery as the use-case of a recommender system, where we consider both relevance and novelty aspects, i.e., recommending music liked by, but previously unknown to users. We propose three new recommendation approaches operating on semantic descriptions, based on the proposed user preference modeling technique. To evaluate them, we compare our methods with two baseline approaches working on metadata. First, we employ a simple approach which uses exclusively genre information for a user's preference examples. Second, we apply a state-of-the-art commercial black-box recommender system on the basis of *Last.fm*. This recommender relies on metadata, and partially uses collaborative filtering information (Levy & Bosteels, 2010), operating on a large database of users and their listening statistics. We provide this system with editorial metadata for the preference examples to retrieve recommendations. Moreover, we also consider two audio content-based baseline approaches. In contrast to the proposed semantic methods, these algorithms use the same procedure for recommendation but operate on low-level timbral features. We then evaluate all considered approaches on 12 subjects, for which we use their gathered preference data to generate recommendations and carry out a listening experiment to assess familiarity, liking and further listening intentions of the provided recommendations. The obtained results indicate that our proposed approaches perform close to metadata-based commercial systems. Moreover, we show that the proposed approaches perform comparably to the baseline approach working on metadata which relies exclusively on manually annotated genre information to represent user preferences and a music collection to recommend music from. Furthermore, the proposed approaches significantly outperformed the low-level timbre-based baselines, supporting our hypothesis on the advantage of using semantic descriptors for music recommendation.

In a second step we exploit the proposed user preference model to map its semantic description to a visual domain. To the best of our knowledge, this task of translating music-oriented user models into visual counterparts has not been explored previously. We propose a novel approach to depict a user's preferences. In our study we consider three descriptor integration methods to represent user preferences in a compact form suitable for mapping it to a visual domain. We evaluate this visualization approach on the same 12 subjects and discuss the obtained results. More precisely, we show that the generated visualizations are able to reflect the subjects' core preferences and are considered by the users as a closely resembling, though not perfect, representation of their musical preferences.

In summary, the proposed technique generates a user model from a set of explicitly provided music tracks, which, in turn, are characterized by the computed semantic descriptors. This semantic representation can be useful in different applications, along with music recommendation, to enrich user experience and increase user trust in a final recommender system. The examples of such applications are, among others, user characterization and visualization, and justification of the provided recommendations. To support and evaluate the proposed technique, we focus on two applications, namely music recommen-

<sup>6</sup> We will use the generic terms “descriptor” and “semantic descriptor” to refer to any high-level semantic description.

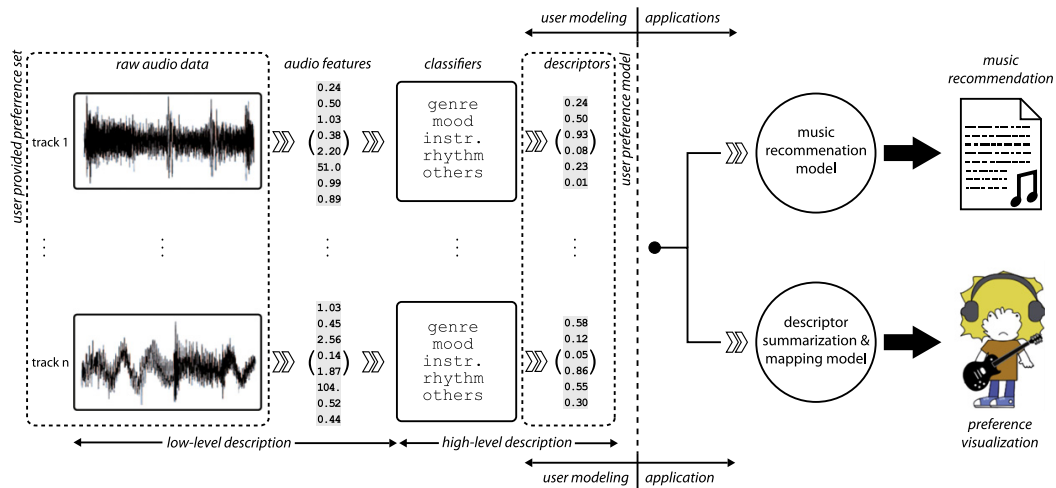


Fig. 1. General scheme of the proposed preference elicitation (user modeling) technique and its applications.

ation and musical preference visualization. A general scheme of the proposed technique and its applications is presented in Fig. 1.

This article is organized as follows: The next section covers related work in the field of audio content-based music recommendation. In Section 3 we describe the proposed preference elicitation technique, including the processes of data gathering Section 3.1 and automatic descriptor extraction Section 3.2. In Section 4 we analyze the evaluation data provided by 12 participants. Section 5 focuses on the first application of the presented preference inference technique – audio content-based music recommendation. In Section 6 we present the second application – audio content-based visualization of musical preferences, starting from the proposed user modeling technique. In Section 7 we provide a general discussion and consider several use-cases of integration of the proposed applications into a final recommender system and their further improvement. Finally, in Section 8 we state general conclusions and highlight future research directions.

## 2. Related work in music recommendation

In this section we review the most important studies in music recommendation, considering both audio content-based and hybrid approaches. These studies can be divided into three categories: personalized music similarity measures, audio content-based models and hybrid models of user preferences.

A number of studies incorporate perceptual personalization of music similarity measures which can be applied for music recommendation. Sotiropoulos et al. (2007) present an active learning system, which adapts the underlying Euclidean distance measure according to a user's feedback on the perceived music similarity. The system operates on sets of timbral, temporal, and tonal features, employing feature selection based on neural networks. Vignoli and Pauws (2005) present a music recommender system based on a hybrid distance measure defined as a user-weighted combination of timbre, genre, tempo, year, and mood distance components. The weights can be explicitly defined by the user. Moreover, Lu and Tseng (2009) present a personalized hybrid recommender system. They propose to combine a distance working on tonal and rhythmic features together with a distance based on collaborative filtering information about preferred tracks, and a semantic emotion-based distance. In order to train the personalized hybrid distance, the user is given a sample of music tracks and is asked to explicitly supply the system with preference assessments (likes/dislikes) and the underlying reasons (such as preference by tonality, and rhythm) for each track. Based on these assessments, the system searches for the closest tracks to the preferred tracks in a music collection using the personalized distance. The scope of this system is considerably limited: its audio content-based component is based on score analysis instead of real audio while the emotion-based component requires manual mood annotations done by experts.

Regarding the work on audio content-based user modeling for music recommendation, Hoashi et al. (2003) present a system with an underlying classification procedure, which divides tracks into the "good" and "bad" categories according to the genre preferences explicitly given by a user. Tree-based vector quantization is used for classification of the tracks represented in a timbral feature space by mel-frequency cepstral coefficients (MFCCs). A sample of tracks labeled by genre is used for initial training of the algorithm. Additional corrections to the classification algorithm can be done via relevance feedback. Grimaldi and Cunningham (2004) apply similar classification using the tracks rated by a user as "good" and "bad" examples. The authors employ kNN and feature sub-space ensemble classifiers working on a set of timbral and temporal features. These classifiers and features were originally suited for the task of genre classification. Due to this fact, the authors found that the

proposed approach fails in the case when the user's preference is not driven by a certain genre. Logan (2004) proposes to generate recommendations based on an explicitly given set of music tracks, which represent a user's preferences. A timbral distance measure is applied to find the tracks similar to the set. As such, the author proposes to use the Earth mover's distance between clusters of MFCCs, which represent music tracks. Unfortunately, no evaluation on real listeners was conducted. Instead, a set of tracks from a randomly chosen album was used to simulate a user's preferences. A track for the same album, not belonging to the user set, is then used as an objective criterion for the evaluation. One of the potential drawbacks of such an evaluation methodology consists in the bias, which leads to the overestimation of real performance, given that timbral distances tend to easily recognize tracks for the same album due to the so-called "album effect" (Mandel & Ellis, 2005). This effect implies that, due to the production process, tracks from the same album share much more timbral characteristics than tracks from different albums of the same artist, and, more so, different artists.

Finally, there are more sophisticated user modeling approaches which use both metadata and audio content information. Yoshii et al. (2006) present a probabilistic user model, which incorporates ratings given by a user and audio content-based "bags-of-timbres". The latter ones represent polyphonic timbre weights, and are obtained from a Gaussian mixture model of MFCCs for each track. The authors use a Bayesian network in the core of their system. A simulation by user ratings obtained from the Amazon Internet store was used to conduct an objective evaluation. Li et al. (2005) and Li, Myaeng, and Kim (2007) propose a track-based probabilistic model, which extends the collaborative filtering approach with audio content-based information. In this model, music tracks are classified into groups based on both available user ratings (by all users in the system) and the extracted set of timbral, rhythmic, and pitch features. The predictions are made based on a user's own ratings, considering their Gaussian distribution on each group of tracks. The authors conducted an objective evaluation using ground truth user ratings. Similarly, Su et al. (2010) present a hybrid recommendation approach, which represents the tracks in a audio content-based feature space. Patterns of temporal evolution of timbral information are computed for each track, represented as frame sequences of clusters of timbral features. Subsequently, given a collaborative filtering information in the form of user ratings, the tracks can be classified into "good" and "bad" according to the ratings of a user and his/her neighbors with similar ratings. To this end, the frequency of the occurrence of "good" and "bad" patterns are computed for each track and are taken as a criterion for classification. The evaluation of the proposed approach is done on ground truth ratings obtained from the Amazon Internet store.

### 3. Methodology

In this section we explain the proposed audio content-based technique for user modeling. We describe the underlying procedure of gathering user preference examples and the process of descriptor extraction. This technique was partially presented in (Bogdanov, Haro, Fuhrmann, Gómez, & Herrera, 2010; Haro et al., 2010).

#### 3.1. Preference examples gathering

As a first step, we ask users to gather the minimal set of music tracks which is sufficient to grasp or convey their musical preferences (the user's *preference set*). Ideally, the selection of representative music should not be biased by any user expectations about a final system or interface design issues. Therefore, for evaluation purposes, we do not inform the user about any further usage of the gathered data, such as giving music recommendations or preference visualization. Furthermore, we do not specify the number of required tracks, leaving this decision to the user.

Generally, example gathering could be performed by either asking the user to provide the selected tracks in audio format (e.g., mp3) or by means of editorial metadata sufficient to reliably identify and retrieve each track (i.e., artist, piece title, edition, etc.). For the proposed audio content-based technique and its applications, the music pieces would be informative even without any additional metadata (such as artist names and track titles). Nevertheless, for a considerable amount of users in a real world (industrial) scenario, providing metadata can be easier than uploading audio. In this case, the audio including full tracks or previews can be obtained from the associated digital libraries by the provided metadata.

For our evaluation purposes only, users are obliged to provide audio files and optionally provide metadata. We then, by means of audio fingerprinting,<sup>7</sup> retrieve and clean metadata for all provided tracks including the ones solely submitted in audio format. Therefore, we will be able to compare our approaches to metadata-based approaches in the case of music recommendation. We also ask the users for additional information, including personal data (gender, age, interest in music, musical background), a description of the strategy followed to select the music pieces, and the way they would describe their musical preferences.

#### 3.2. Descriptor extraction

Here we describe the procedure followed to obtain a semantic representation of each music track from the user's preference set. We follow Bogdanov et al. (2009) and Bogdanov, Serrà, Wack, Herrera, and Serra (2011) to obtain such descriptions.

<sup>7</sup> We use MusicBrainz service: [http://musicbrainz.org/doc/MusicBrainz\\_Picard](http://musicbrainz.org/doc/MusicBrainz_Picard).

**Table 1**

Ground truth music collections employed for semantic regression. Source references: (1) Homburg et al. (2005), (2) in-house, (3) Tzanetakis and Cook (2002), (4) Gómez and Herrera (2008), (5) Laurier et al. (2009) + in-house, and (6) Cano et al. (2006).

Name	Category	Classes (semantic descriptors)	Size (tracks)	Source
G1	Genre & Culture	Alternative, blues, electronic, folk/country, funk/soul/rnb, jazz, pop, rap/hiphop, rock	1820 track excerpts, 46–490 per genre	(1)
G2	Genre & Culture	Classical, dance, hip-hop, jazz, pop, rhythm'n'blues, rock, speech	400 tracks, 50 per genre	(2)
G3	Genre & Culture	Blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, rock	993 track excerpts, 100 per genre	(3)
CUL	Genre & Culture	Western, non-western	1640 track excerpts, 1132/508 per class	(4)
MHA	Moods & Instruments	Happy, non-happy	302 full tracks + excerpts, 139/163 per class	(5)
MSA	Moods & Instruments	Sad, non-sad	230 full tracks + excerpts, 96/134 per class	(5)
MAG	Moods & Instruments	Aggressive, non-aggressive	280 full tracks + excerpts, 133/147 per class	(5)
MRE	Moods & Instruments	Relaxed, non-relaxed	446 full tracks + excerpts, 145/301 per class	(5)
MAC	Moods & Instruments	Acoustic, non-acoustic	321 full tracks + excerpts, 193/128 per class	(5)
MEL	Moods & Instruments	Electronic, non-electronic	332 full tracks + excerpts, 164/168 per class	(5)
RPS	Rhythm & Tempo	Perceptual speed: slow, medium, fast	3000 full tracks, 1000 per class	(2)
RBL	Rhythm & Tempo	Chachacha, jive, quickstep, rumba, samba, tango, viennese waltz, waltz	683 track excerpts, 60–110 per class	(6)
ODA	Other	Danceable, non-danceable	306 full tracks, 124/182 per class	(2)
OPA	Other	Party, non-party	349 full tracks + excerpts, 198/151 per class	(2)
OVI	Other	Voice, instrumental	1000 track excerpts, 500 per class	(2)
OTN	Other	Tonal, atonal	345 track excerpts, 200/145 per class	(2)
OTB	Other	Timbre: bright, dark	3000 track excerpts, 1000 per class	(2)

For each music track, we calculate a low-level feature representation using an in-house audio analysis tool.<sup>8</sup> In total, this tool provides over 60 commonly used low-level audio features, characterizing global properties of the given tracks, related to timbral, temporal, and tonal information. The features include inharmonicity, odd-to-even harmonic energy ratio, tristimuli, spectral centroid, spread, skewness, kurtosis, decrease, flatness, crest, and roll-off factors, MFCCs, spectral energy bands, zero-crossing rate (Peeters, 2004), spectral complexity (Streich, 2007), transposed and untransposed harmonic pitch class profiles, key strength, tuning, chords (Gómez, 2006), pitch, beats per minute (BPM) and onsets (Brossier, 2007). Most of these features are extracted on a frame-by-frame basis and then summarized by their means and variances across all frames. In the case of multidimensional features (e.g., MFCCs), covariances between components are also considered.

We use the described low-level features to infer semantic descriptors. To this end, we perform a regression by suitably trained classifiers producing different semantic dimensions such as genre, musical culture, moods, instrumentation, rhythm, and tempo. We opt for multi-class support vector machines (SVMs) with a one-vs.-one voting strategy (Bishop, 2006), and use the libSVM implementation.<sup>9</sup> In addition to simple classification, this implementation extends the capabilities of SVMs making available class probability estimation (Chang & Lin, 2011), which is based on the improved algorithm by Platt (2000). The classifiers are trained on 17 ground truth music collections (including full tracks and excerpts) presented in Table 1, corresponding to 17 classification tasks.

For each given track, each classifier returns the probabilistic estimates of classes on which it was trained. The classifiers operate on optimized low-level feature representations of tracks. More concretely, each classifier is trained on a reduced set of features, which is individually selected based on correlation-based feature selection (Hall, 2000) according to the underlying music collection. Moreover, the parameters of each SVM are found by a grid search with 5-fold cross-validation. Classification results form a high-level semantic descriptor space, which contains the probability estimates for each class of each classifier. The accuracy of classifiers varies between 60.3% and 98.2% with the median accuracy being 88.2%. Classifiers trained on G1 and RBL show the worst performance, close to 60%,<sup>10</sup> while classifiers for CUL, MAG, MRE, MAC, OVI, and OTB show the best performance, greater than 93%.

With the described procedure we obtain 62 semantic descriptors, shown in Table 1, for each track in the user's preference set. These resulting representations of tracks (i.e., vectors of class probabilities) form our proposed user model, defined as a set  $U$ :

$$U = \{P(C_{1,1}|T_i), \dots, P(C_{1,N_1}|T_i), \dots, P(C_{17,1}|T_i), \dots, P(C_{17,N_{17}}|T_i)\}, \quad (1)$$

<sup>8</sup> <http://mtg.upf.edu/technologies/essentia>.

<sup>9</sup> <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>.

<sup>10</sup> Still, note the amount of classes in G1 and RBL classifiers is 9 and 3, respectively.

where  $P(C_{k,l}|T_i)$  stands for the probability of track  $T_i$  from a preference set belonging of  $l$ th class  $C_{k,l}$  of the  $k$ th classifier having  $N_k$  classes.

As the procedure of the low-level signal analysis and the details of semantic descriptor extraction are out of the scope of this paper, we refer the interested reader to the aforementioned literature on low-level features, and to (Bogdanov et al., 2009, 2011), and references therein, for details on the SVM implementation.

#### 4. User data analysis

In order to evaluate the proposed technique, we worked with a group of 12 participants (8 male and 4 female) selected from the authors' colleagues and acquaintances without disclosing any detail of the targeted research. They were aged between 25 and 45 years old (average  $\mu = 33$  and standard deviation  $\sigma = 5.35$ ) and showed a very high interest in music (rating around  $\mu = 9.64$ , with  $\sigma = 0.67$ , where 0 means no interest in music and 10 means passionate about music). Ten of the 12 participants play at least one musical instrument, including violin, piano, guitar, synthesizers, and ukulele.

The number of tracks selected by the participants to convey their musical preferences was very varied, ranging from 23 to 178 music pieces ( $\mu = 73.58$ ,  $\sigma = 45.66$ ) with the median being 57 tracks. The time spent for this task also differed a lot, ranging from half an hour to 60 h ( $\mu = 11.11$ ,  $\sigma = 22.24$ ) with the median being 5 h.

It is interesting to analyze the provided verbal descriptions about the strategy followed to select the music tracks. Some of the participants were selecting one track per artist, while some others did not apply this restriction. They also covered various uses of music such as listening, playing, singing or dancing. Other participants mentioned musical genre, mood, expressivity, musical qualities, and chronological order as driving criteria for selecting the tracks. Furthermore, some participants implemented an iterative procedure by gathering a very large amount of music pieces from their music collections and performing a further refinement to obtain the final selection. Finally, all participants provided a set of labels to define their musical preferences. We asked them to provide labels related to the following aspects: musical genre, mood, instrumentation, rhythm, melody/harmony, and musical expression. We also included a free category for additional labels on top of the proposed musical facets.

The number of labels provided by the participants ranged from 4 to 94 labels ( $\mu = 25.11$ ,  $\sigma = 23.82$ ). The distribution of the number of labels that participants provided for each facet (normalized by the total number of labels provided by each participant) is presented in Fig. 2. We observe that most of them were related to genre, mood, and instrumentation, some of them to rhythm and few to melody, harmony, or musical expression. Other suggested labels were related to lyrics, year, and duration of the piece. The participants' preferences covered a wide range of musical styles (e.g., classical, country, jazz, rock, pop, electronic, folk), historical periods, and musical properties (e.g., acoustic vs. synthetic, calm vs. danceable, tonal vs. atonal). Taking into account this information, we consider that the population represented by our participants corresponds to that of music enthusiasts, but not necessarily mainstream music consumers.

Finally, the music provided by the participants was very diverse. Fig. 3 presents an overall tag cloud of music preferences of our population (mostly genre-based). The tag cloud was generated using artist tags found on *Last.fm* tagging service for all tracks provided by the participants with a normalization by the number of tracks provided by each participant.

#### 5. Music recommendation

The first considered application exploits the computed user model to generate music recommendations based on semantic descriptors. For consistency, we focus on the task of retrieving 20 music tracks from a given music collection as recom-

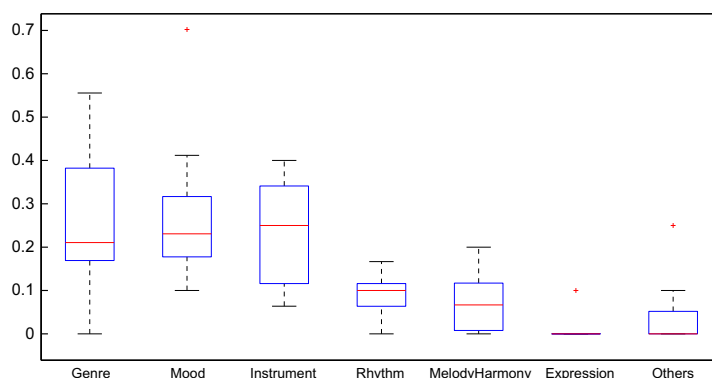


Fig. 2. Box plot of the proportions of provided labels per musical facet, normalized by the total number of labels per participant. Categories from left to right correspond to genre, moods, instruments, rhythm, melody and harmony, musical expression, and other labels respectively. Red crosses stand for extreme outliers. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



2. *Semantic distance from all tracks (SEM-ALL)*. Alternatively, instead of simplifying the user model to one point we consider all individual tracks. Thus, we take into account all possible areas of preferences, explicitly specified by the user, while searching for the most similar tracks. We define a track-to-set semantic distance as a minimum semantic distance from a track to any of the tracks in the preference set. We return the 20 nearest tracks according to this distance as recommendations.
3. *Semantic Gaussian mixture model (SEM-GMM)*. Finally, we propose to represent the user model as a probability density of preferences in the semantic space. We employ a Gaussian mixture model (GMM) (Bishop, 2006), which estimates a probability density as a weighted sum of a given number of simple Gaussian densities (components). The GMM is initialized by k-mean clustering, and is trained with an expectation–maximization algorithm. We select the number of components in the range between 1 and 20, using a Bayesian information criterion (Bishop, 2006). Once we have trained the model, we compute the probability density for each of the tracks. We rank the tracks according to the obtained density values<sup>11</sup> and return the 20 most probable tracks as recommendations.

## 5.2. Evaluation

Here we describe the evaluation of the proposed recommendation approaches against metadata-based and audio content-based baselines.

### 5.2.1. Metadata-based baselines

We consider two baseline approaches to music recommendation working on metadata. The first baseline is constructed exclusively using information about the user's genre preferences. The second one is based on the information about preferred tracks and artists (taken from the editorial metadata provided by the user for the preference set), and partially employs collaborative filtering information, querying a commercial state-of-the-art music recommender for similar music tracks.

1. *Random tracks from the same genre (GENRE)*. This simple and low-cost approach provides random recommendations relying on genre categories of the user's preference set. We assume that all tracks in the given music collection are manually tagged with a genre category by an expert. We randomly preselect 20 tracks from the preference set and obtain their genre labels. Ideally, tracks from the preference set should contain manual genre annotations by an expert as well. Moreover, the annotations should be consistent with the ones in the music collection to be able to match the tracks by genre. Nevertheless, the tracks from the preference set, since they were submitted by the user, do not necessarily contain a genre tag, and the quality of such tags and their consistency with the genres in the music collection cannot be assured. Therefore, we retrieve this information from the Web. We use track pages or artist pages from the social music tagging system *Last.fm* as the source of genre information. We run queries using metadata of the preselected tracks, and select the most popular genre tag, which is presented among genre tags of the given music collection. For each of the 20 preselected tracks, we return a random track of the same genre label.
2. *Black-box music similarity from Last.fm (LASTFM)*. As we did not have collaborative filtering data available for our research (and moreover, a large dataset would be required to match with our participants' tracks), we opted to use black box recommendations provided by *Last.fm*.<sup>12</sup> It is an established music recommender with an extensive number of users, and a large playable music collection, providing means for both monitoring listening statistics and social tagging (Jones & Pu, 2007). In particular, it provides track-to-track<sup>13</sup> and artist-to-artist<sup>14</sup> similarity computed by the undisclosed algorithm, which is partially based on collaborative filtering, but does not use any audio content. It is important to notice that the underlying music collection of *Last.fm* used in this baseline approach differs (being significantly larger and broader) from the collection used by the other approaches in our evaluation. Again, we randomly preselect 20 tracks from the preference set and independently query *Last.fm* for each of them to receive a recommendation. For each track we select the most similar track from the recommended ones with an available preview.<sup>15</sup> If no track-based similarity information is available (e.g., when the query track is an unpopular long-tail track with a low number of listeners), we query for similar artists. In this case we choose the most similar artist and select its most popular track with an available preview.

### 5.2.2. Audio content-based baselines

We consider two audio content-based baseline approaches. These approaches apply the same ideas as the proposed semantic approaches, but operate on low-level timbral features, frequently used in the related literature.

1. *Timbral distance from all tracks (MFCC-ALL)*. This approach is a counterpart to the proposed *SEM-ALL* approach using a common low-level timbral distance (Pampalk, 2006) instead of the semantic one. The tracks are modeled by probability dis-

<sup>11</sup> Under the assumption of a uniform distribution of the tracks in the universe within the semantic space.

<sup>12</sup> All experiments were conducted on May 2010.

<sup>13</sup> For example, [http://last.fm/music/Grandmaster+Flash/\\_/The+Message/+similar](http://last.fm/music/Grandmaster+Flash/_/The+Message/+similar).

<sup>14</sup> For example, <http://last.fm/music/Baby+Ford/+similar>.

<sup>15</sup> These previews are downloadable music excerpts (30 s), which are later used in our subjective evaluation for the case of the *LASTFM* approach.

tributions of MFCCs using single Gaussian with full covariance matrix. For such representations a distance measure can be defined using a closed form approximation of the Kullback–Leibler divergence. This baseline resembles the state-of-the-art timbral user model, proposed by Logan (2004), which uses the Earth-Mover’s Distance between MFCC distributions as a distance.

2. *Timbral Gaussian mixture model (MFCC-GMM)*. Alternatively, we consider a counterpart to the proposed *SEM-GMM* probabilistic approach: we use a population of mean MFCC vectors (one vector per track from the user’s preference set) to train a timbral GMM.

### 5.2.3. Evaluation methodology

We performed subjective listening tests on our 12 subjects in order to evaluate the considered approaches. As the source for recommendations, we employed a large in-house music collection, covering a wide range of genres, styles, arrangements, geographic locations, and musical epochs. This collection consists of 100,000 music excerpts (30 s) by 47,000 artists (approximately 2 tracks per artist).

For each subject, we computed the user model from the provided preference set. According to the considered recommendation approaches we generated 7 playlists (three by the proposed approaches working with the semantic user model, two by the approaches working on metadata, and two by the low-level timbral approaches). Each playlist consisted of 20 music tracks. Following a usual procedure for evaluation of music similarity measures and music recommendations, we applied an artist filter (Pampalk, 2006) to assure that no playlist contained more than one track from the same artist nor tracks by the artists from the preference set. These playlists were merged into a single list, in which tracks were randomly ordered and anonymized, including filenames and metadata. The tracks offered as recommendations were equally likely to come from each single recommendation approach. This allowed us to avoid any response bias due to presentation order, recommendation approach, or contextual recognition of tracks (by artist names, etc.) by the participants. In addition, the participants were not aware of the amount of recommendation approaches, their names and their rationales.

We designed a questionnaire in order to obtain the different subjective impressions related to the recommended music (see Table 2). For each recommended track the participants were asked to provide a number of ratings:

- *Familiarity* ranged from 0 to 4; with 0 meaning absolute unfamiliarity, 1 feeling familiar with the music, 2 knowing the artist, 3 knowing the title, and 4 the identification of artist and title.
- *Liking* measured the enjoyment of the presented music with 0 and 1 covering negative liking, 2 representing a neutral position, and 3 and 4 representing increasing liking for the musical excerpt.
- *Listening intentions* measured the readiness of the participant to listen to the same track again in the future. This measure is more direct and behavioral than the *liking*, as an intention is closer to action than just the abstraction of liking. Again the scale contained 2 positive and 2 negative steps plus a neutral one.
- “*Give-me-more*” with 1 indicating request for more music like the presented track, and 0 indicating reject of such music.

The users were also asked to provide the track title and artist name for those tracks rated high in the familiarity scale.

### 5.2.4. Results

First, we manually corrected familiarity ratings when the artist/title provided by a user was incorrect compared to the actual ones. In such situations, a familiarity rating of 3, or, more frequently, 4 or 2, was lowered to 1 (in the case of incorrect

**Table 2**  
Meaning of familiarity, liking, listening intentions, and “give-me-more” ratings as given to the participants.

Rating	Value	Meaning
Familiarity	4	I know the song and the artist
	3	I know the song but not the artist
	2	I know the artist but not the song
	1	It sounds familiar to me even I ignore the title and artist (maybe I heard it in TV, in a soundtrack, long time ago, etc.)
	0	No idea
Liking	4	I like it a lot!
	3	I like it
	2	I would not say I like it, but it is listenable
	1	I do not like it
	0	It is annoying, I cannot listen to it!
Listening intentions	4	I am going to play it again several times in the future
	3	I probably will play it again in the future
	2	It does not annoy me listening to it, although I am not sure about playing it again in the future
	1	I am not going to play it again in the future
Give-me-more	0	I will skip it in any occasion I find in a playlist
	1	I would like to be recommended more songs like this one
	0	I would not like to be recommended more songs like this one



artist and track title) or 2 (in the case of correct artist, but incorrect track title). These corrections represented just 3% of the total familiarity judgments.

Considering the subjective scales used, a good recommender system should provide high-liking/listening intentions/request for the greater part of retrieved tracks and in particular for low-familiarity tracks. Therefore, we recoded the user's ratings into 3 main categories, referring to the type of the recommendation: *hits*, *fails*, and *trusts*. Hits were those tracks having a low familiarity rating (<2), high (>2) liking and intentions ratings, and a positive (>0) "give-me-more" request. Fails were those tracks having low (<3) liking and intentions ratings, and null "give-me-more" request. Trusts were those tracks which got a high familiarity (>1), high (>2) liking and intentions ratings, and a positive (>0) "give-me-more" request. Trusts, provided their overall amount is low, can be useful for a user to feel that the recommender is understanding his/her preferences (Barrington et al., 2009; Cramer et al., 2008). A user could be satisfied by getting a trust track from time to time, but annoyed if every other track is a trust, especially in the use-case of music discovery (the main focus of the present work). 18.3% of all the recommendations were considered as "unclear" (e.g., a case when a track received a high liking, but a low intentions rating and a null "give-me-more" request). Most of the unclear recommendations (41.9%) consisted of low liking and intention ratings (<3 in both cases) followed by a positive "give-me-more" request; other frequent cases of unclear recommendation consisted of a positive liking (>2) that was not followed by positive intentions and positive "give-me-more" (15.5%) or positive liking not followed by positive intentions though positive "give-me-more" (20.0%). We excluded the unclear recommendations from further analysis.

We report the percent of each outcome category per recommendation approach in Table 3 and Fig. 5a. An inspection of it reveals that the approach which yields the largest amount of hits (41.2%) and trusts (25.4%) is *LASTFM*. The trusts found with other approaches were scarce, all below 4%. The approaches based on the proposed semantic user model (*SEM-ALL*, *SEM-MEAN* and *SEM-GMM*) yielded more than 30% of hits, and the remaining ones did not surpass 25%. The existence of an association between recommendation approach and the outcome of the recommendation was statistically significant, according to the result of the Pearson chi-square test ( $\chi^2(18) = 351.7, p < 0.001$ ).

Additionally, we performed three separate between-subjects ANOVA tests in order to test the effects of the recommendation approaches on the liking, intentions, and "give-me-more" subjective ratings. The effect was confirmed in all of them ( $F(6, 1365) = 55.385, p < 0.001$  for the liking rating,  $F(6, 1365) = 48.89, p < 0.001$  for the intentions rating, and  $F(6, 1365) = 43.501, p < 0.001$  for the "give-me-more" rating). Pairwise comparisons using Tukey's test revealed the same pattern of differences between the recommendation approaches, irrespective of the three tested indexes. This pattern highlights the *LASTFM* approach as the one getting the highest overall ratings. It also groups together the timbral *MFCC-GMM* and *MFCC-ALL* approaches (those getting the lowest ratings), and the remaining approaches (*SEM-ALL*, *SEM-MEAN*, *SEM-GMM*, and *GENRE*) are grouped in-between. The mean values of the obtained liking, listening intentions, and "give-me-more" ratings per each approach are presented in Fig. 5b.

Finally, a measure of the quality of the hits was computed by multiplying the difference of liking and familiarity by listening intentions for each recommended track. This quality score ranks recommendations considering that the best ones correspond to the tracks which are highly-liked though completely unfamiliar, and intended to be listened again. Selecting only the hits, an ANOVA on the effect of the recommendation approach on this quality measure revealed no significant differences between any of the approaches. Therefore, considering the quality of hits, there is no recommendation approach granting better or worst recommendations than any other. The same pattern was revealed by solely using the liking as a measure of the quality of the hits.

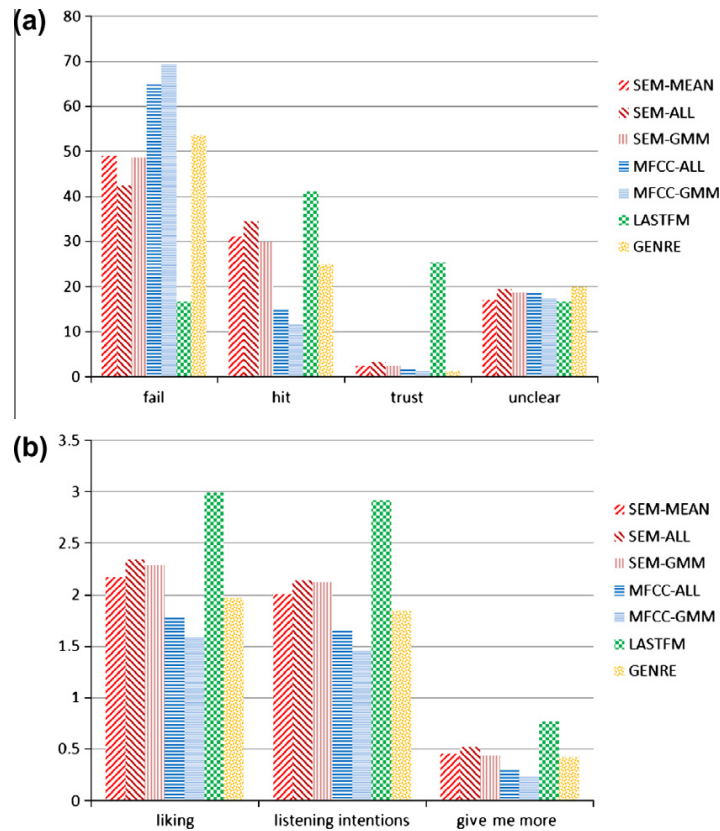
### 5.3. Discussion

We presented an application of the considered user model for music recommendation. Based on this computed model, we proposed three approaches operating on a subset of the retrieved semantic descriptors. Two of these approaches recommend tracks similar to the preference set using a semantic distance. The third approach creates a probabilistic model using GMM to estimate the density of the user's preferences within the semantic domain. We evaluated these approaches against two metadata-based and two audio content-based baselines in a subjective evaluation on 12 participants. Specifically, we employed a simple metadata-based approach which recommends random tracks, selected from the genres preferred by the

**Table 3**

The percent of fail, trust, hit, and unclear categories per recommendation approach. Note that the results for the *LASTFM* approach were obtained on a different underlying music collection.

Approach	Fail	Hit	Trust	Unclear
<i>SEM-MEAN</i>	49.167	31.250	2.500	17.083
<i>SEM-ALL</i>	42.500	34.583	3.333	19.583
<i>SEM-GMM</i>	48.750	30.000	2.500	18.750
<i>MFCC-ALL</i>	64.167	15.000	2.083	18.750
<i>MFCC-GMM</i>	69.583	11.667	1.250	17.500
<i>LASTFM</i>	16.667	41.250	25.417	16.667
<i>GENRE</i>	53.750	25.000	1.250	20.000



**Fig. 5.** The percent of fail, trust, hits, and clear categories per recommendation approach (a); the liking, listening intentions, and “give-me-more” mean ratings for recommendation approach (b). The results for the *LASTFM* approach were obtained on a different underlying music collection. The “give-me-more” rating varies in the  $[0,1]$  interval.

user. Alternatively, given the editorial metadata for the user’s preference set, we employed a state-of-the-art black-box recommender working on collaborative filtering information – *Last.fm*, to retrieve similar music. Among the audio content-based baselines, we employed two approaches operating on low-level timbral features (MFCCs) instead of the semantic descriptors. These approaches are counterparts to our semantic distance-based and probabilistic approaches, working with a timbral user model.

The evaluation results revealed the users’ preference for the proposed semantic approaches over the low-level timbral baselines. This fact supports our hypothesis on the advantage of using semantic description for music recommendation. Moreover, it complements the outcomes from the previous research on semantic music similarity (Bogdanov et al., 2009). We may conclude that the high-level semantic description outperforms the low-level timbral description in the task of music recommendation and, in particular, musical preference elicitation.

Comparing with the baselines working on metadata, we found that the proposed approaches perform better than the simple genre-based recommender (although no statistically significant differences were found in terms of liking, listening intentions, and “give-me-more” ratings). Interestingly, this naive genre-based recommender still outperformed the timbre-based baselines. This could be partially explained by the fact that genre was one of the driving criteria for selecting the users’ preference sets (see Fig. 2), and that manually annotated genre and sub-genre labels entail more information and diversity than timbral information automatically extracted from MFCCs.

On the other hand, the proposed approaches were found to be inferior to the considered commercial recommender (*LASTFM*) in terms of the number of successful novel recommendations (hits). Still, this metadata-based approach using collaborative filtering yielded only 7 absolute percentage points more hits than one of our proposed semantic methods (*SEM-ALL*). Considering trusted recommendations, the *LASTFM* baseline provided about 22% more recommendations already known by the participants. Interestingly, one track out of four recommended by this baseline was already familiar to the participants, which might be considered an excessive amount considering the music discovery use-case. In particular, the larger amount

of both hits and trusts provided by the *LASTFM* baseline can be partly explained by the fact that the recommendations were generated using the *Last.fm* music collection. Due to the extensive size of this collection and the large amount of available collaborative filtering data, we can hypothesize the obtained performance of this approach to be an upper bound in both hits and trusts and expect a lower performance on our smaller in-house collection. Taking all this into account, we expect the proposed semantic approaches, and the underlying semantic user model, to be suitable for music discovery in the long tail which can suffer from insufficient, incorrect, or incomplete metadata information.

## 6. Visualization of musical preferences

The second application exploits the computed user model to generate a visualization of the user's musical preferences in form of a *Musical Avatar*, a humanoid cartoon-like character. Although such a task is not directly related to music recommendation, it might be a useful enhancement for recommender systems. In particular, automatic user visualization can provide means to increase user engagement in the system, justify recommendations (e.g., by visualizing playlists), and facilitate social interaction among users.

### 6.1. Descriptor summarization

The retrieved semantic descriptors provide a rich representation of user preferences, which in particular can give valuable cues for visualization. Instead of using their full potential, in this proof-of-concept application we operate on a reduced subset of descriptors for simplicity reasons in the mapping process. To this end, we select this subset considering the classifiers' accuracy against ground truth values provided by a subset of five participants. When selecting the subset, we also intend to preserve the representativeness of the semantic space. We asked these participants to manually annotate their own music collections with the same semantic descriptors as those inferred by the classifiers. We then compared these manual annotations with the classifiers' outputs by Pearson correlation and selected the best performing descriptors. The observed correlation values for all semantic descriptors varied between  $-0.05$  and  $0.70$  with the median being  $0.40$ . The subset of 17 descriptors was selected with the majority of correlations (for 14 descriptors) being greater than  $0.40$ . The resulting descriptors, which are used by the proposed visualization approach, are presented in Table 4.

Having refined the semantic descriptors for the computed user model, we consider different summarization methods to obtain a compact representation which can be mapped to the visual domain. With these summarization strategies we explore the degree of descriptor resolution necessary for optimal visual representation. These strategies can be based on continuous or discrete values, and therefore lead to visual elements of continuous or discrete nature (e.g., size). The idea behind this exploration is related to the possibility that users might prefer simpler objects (discrete visual elements such as presence or absence of a guitar) or more complex ones (continuous elements such as guitars of different sizes) depicting subtle variations of preferences.

We summarize the user model across individual tracks to a single multidimensional point in a semantic descriptor space as in the case of the *SEM-MEAN* representation proposed for music recommendation (Section 5.1). We first standardize each descriptor to remove global scaling and spread; i.e., for each track from the user's preference set we subtract the global mean and divide by the global standard deviation. We estimate the reference means ( $\mu_{R,i}$ ) and standard deviations ( $\sigma_{R,i}$ ) for each descriptor from the representative in-house music collection of 100,000 music excerpts used for the subjective evaluation of music recommendation approaches (Section 5.2.3). Moreover, we range-normalize the aforementioned standardized descriptor values according to the following equation:

$$N_i = \frac{d_i - \min}{\max - \min}, \quad (2)$$

where  $d_i$  is the standardized value of descriptor  $i$ , and since  $d_i$  has zero mean and unit variance, we set the respective *min* and *max* values to  $-3$  and  $3$ , since according to Chebyshev's inequality at least 89 % of the data lies within 3 standard deviations from its mean value (Grimmett & Stirzaker, 2001). We clip all resulting values smaller than 0 or greater than 1. The obtained scale can be seen as a measure of preference for a given category, and is used by the visualization process (see Section 6.2). We then summarize the descriptor values across tracks by computing the mean for every normalized descriptor ( $\mu_{N,i}$ ).

**Table 4**  
Selected descriptors, and the corresponding music collections used for regression, per category of semantic descriptors (i.e., genre, moods & instruments, and others) used for visualization.

Genre	Moods & Instruments	Others
Electronic (G1)	Happy (MHA)	Party (OPA)
Dance (G2)	Sad (MSA)	Vocal (OVI)
Rock (G2)	Aggressive (MAG)	Tonal (OTN)
Classical (G3)	Relaxed (MRE)	Bright (OTB)
Jazz (G3)	Electronic (MEL)	Danceable (ODA)
Metal (G3)	Acoustic (MAC)	

At this point, we consider three different methods to quantize the obtained mean values. These quantization methods convey different degrees of data variability, and are defined as follows:

- *Binary* forces the descriptors to be either 1 or 0, representing only two levels of preference (i.e., 100% or 0%). We quantize all  $\mu_{N_i}$  values below 0.5 to zero and all values above (or equal) 0.5 to one.
- *Ternary* introduces a third value representing a neutral degree of preference (i.e., 50%). We perform the quantization directly from the original descriptor values, that is, we calculate the mean values for every descriptor ( $\mu_i$ ) and quantize them according to the following criteria:

$$Ternary_i = \begin{cases} 1 & \text{if } \mu_i > (\mu_{R_i} + th_i), \\ 0.5 & \text{if } (\mu_{R_i} - th_i) \leq \mu_i \leq (\mu_{R_i} + th_i), \\ 0 & \text{if } \mu_i < (\mu_{R_i} - th_i), \end{cases} \quad (3)$$

where  $th_i = \sigma_{R_i}/3$ .

- *Continuous* preserves all possible degrees of preference. We maintain the computed  $\mu_{N_i}$  values without further changes.

At the end of this process we obtain three simplified representations of the user model, each of them consisting of 17 semantic descriptors.

## 6.2. Visualization

In order to generate the *Musical Avatar*, we convert the summarized semantic descriptors to a set of visual features. According to MacDonald, Hargreaves, and Miell (2002), individual, cultural and sub-cultural musical identities emerge through social groups concerning different types of moods, behaviors, values or attitudes. We apply the cultural approach of representing urban tribes (Maffesoli, 1996), since in these tribes, or subcultures, music plays a relevant role in both personal and cultural identities. Moreover, they are often identified by specific symbolisms which can be recognized visually.

Therefore, we decided to map the semantic descriptors into a basic collection of cultural symbols. As a proof-of-concept, we opt for an iconic cartoon style of visualization. This choice is supported by a number of reasons; firstly, this style is a less time-consuming technique compared to other approaches more focused on realistic features (Ahmed, de Aguiar, Theobalt, Magnor, & Seidel, 2005; Petajan, 2005; Sauer & Yang, 2009). Secondly, it is a graphical medium which, by eliminating superfluous features, amplifies the remaining characteristics of a personality (McCloud, 2009). Thirdly, there are examples of existing popular avatar collections of this kind such as Meegos<sup>16</sup> or Yahoo Avatars.<sup>17</sup>

In our approach the relevant role is played by the graphical symbols, which are filled with arbitrary colors related to them. Although colors have been successfully associated with musical genres (Holm, Aaltonen, & Siirtola, 2009) or moods (Voong & Beale, 2007), the disadvantage of using only colors is the difficulty to establish a global mapping due to reported cultural differences about their meaning.

In our design, we consider the information provided by the selected descriptors and the design requirements of modularity and autonomy. Starting from a neutral character,<sup>18</sup> we divide the body into different parts (e.g., head, eyes, mouth). For each of the parts we define a set of groups of graphic symbols (graphic groups) to be mapped with certain descriptors. Each of these graphic groups always refers to the same set of descriptors. For example, the graphic group corresponding to the mouth is always defined by the descriptors from the categories “Moods and Instruments” and “Others” but never from “Genre” category. The relation between graphic groups and categories of the semantic descriptors is presented in Table 5. For this mapping, we consider the feasibility of representing the descriptors (e.g., the suit graphic group is more likely to represent a musical genre compared to the other descriptor categories). We also bear in mind a proportional distribution between the three main descriptor categories vs. each of these graphic groups in order to notice them all. However, in accordance with the cartoon style some of these graphic groups refer to all three main descriptor categories because they can highlight better the most prominent characteristics of the user's profile, and also they can represent a wide range of descriptors (e.g., the head and complement graphic groups). Apart from the listed graphic groups, we introduce a label to identify the gender of the avatar, each providing a unique set of graphic symbols.

Besides the body elements, we also add a set of possible backgrounds to the graphic collection in order to support some descriptors of the “Others” category such as “party”, “tonal”, or “danceable”. In addition, the “bright” descriptor is mapped to a gray background color that ranges from RGB (100,100,100) to RGB (200,200,200). The relation between graphic groups and categories of the semantic descriptors is presented in Table 5. We note that our decisions on the design, and in particular on the descriptor mapping, are arbitrary, being a matter of choice, of visual and graphic sense, and common sense according to many urban styles of self-imaging.

<sup>16</sup> <http://meegos.com>.

<sup>17</sup> <http://avatars.yahoo.com>.





<sup>18</sup> A neutral character corresponds to an empty avatar. It should be noted that the same representation can be achieved if all normalized descriptor values are set to 0.5 meaning no preference to any descriptor at all.

**Table 5**  
Mapping of the descriptor categories to the graphic groups.

Graphic group	Descriptor categories		
	Genre	Moods & Inst.	Others
Background			•
Head	•	•	•
Eyes		•	•
Mouth		•	•
Complement	•	•	•
Suit	•		•
Hair	•		
Hat	•	•	
Complement2			•
Instrument	•	•	

**Table 6**

Vector representation example: user profile vs. the instrument graphic group (continuous summarization). A visual element with the minimum distance to the user profile is selected (in this case, the turntable).

Category	Descriptor	User profile				
Genre	Classical (G3)	0.0	0.0	0.0	0.0	0.0
Genre	Electronic (G1)	1.0	0.0	0.0	0.0	1.0
Genre	Jazz (G3)	0.0	0.0	1.0	0.0	0.0
Genre	Metal (G3)	0.0	0.0	0.0	0.0	0.0
Genre	Dance (G2)	1.0	0.0	0.0	0.0	0.0
Genre	Rock (G2)	0.5	1.0	0.0	0.0	0.0
Moods & Inst.	Electronic (MEL)	1.0	0.0	0.0	0.0	1.0
Moods & Inst.	Relaxed (MRE)	0.0	0.0	0.0	0.0	0.0
Moods & Inst.	Acoustic (MAC)	0.8	0.0	0.0	1.0	0.0
Moods & Inst.	Sad (MSA)	0.0	0.0	0.0	0.0	0.0
Moods & Inst.	Aggressive (MAG)	0.0	1.0	0.0	0.0	0.0
Moods & Inst.	Happy (MHA)	1.0	0.0	0.0	0.0	0.0
Distance to user profile			2.43	2.62	2.07	<b>1.70</b>

We construct a vector space model and use a Euclidean distance as a measure of dissimilarity to represent the user's musical preferences in terms of graphic elements. For each graphic group we choose the best graphic symbol among the set of all available candidates, i.e., the closest to the corresponding subset of the user's vector model (see Table 6 for an example of the vector representation of these elements). This subset is defined according to the mapping criteria depicted in Table 5. As a result, a particular *Musical Avatar* is generated for the user's musical preferences. All graphics are done in vector format for rescalability and implemented using Processing<sup>19</sup> (Reas & Fry, 2007).

According to the summarization methods considered in Section 6.1, the mapping is done from either a discrete or continuous space resulting in different data interpretations and visual outputs. These differences imply that in some cases the graphic symbols have to be defined differently. For instance, the “vocal” descriptor set to 0.5 in the case of *continuous* method means “she likes both instrumental and vocal music”, while this neutrality is not present in the case of the *binary* method. Furthermore, in the *continuous* method, properties such as size or chromatic gamma of the graphic symbols are exploited while this is not possible within the discrete vector spaces. Fig. 6 shows a graphical example of our visualization strategy where, given the summarized binary user model, the best graphic symbol for each graphic group is chosen. Fig. 7 shows a sample of *Musical Avatars* generated by the three summarization methods and Fig. 8 shows a random sample of different *Musical Avatars*.

### 6.3. Evaluation

#### 6.3.1. Evaluation methodology

We carried out a subjective evaluation on our 12 subjects. For each participant, we generated three *Musical Avatars* corresponding to the three considered summarization methods. We then asked the participants to answer a brief evaluation questionnaire. The evaluation consisted in performing the following two tasks.

In the first task, we asked the participants to manually assign values for the 17 semantic descriptors used to summarize their musical preferences (see Table 4). We requested a real number between 0 and 1 to rate the degree of preference for

<sup>19</sup> <http://processing.org>.

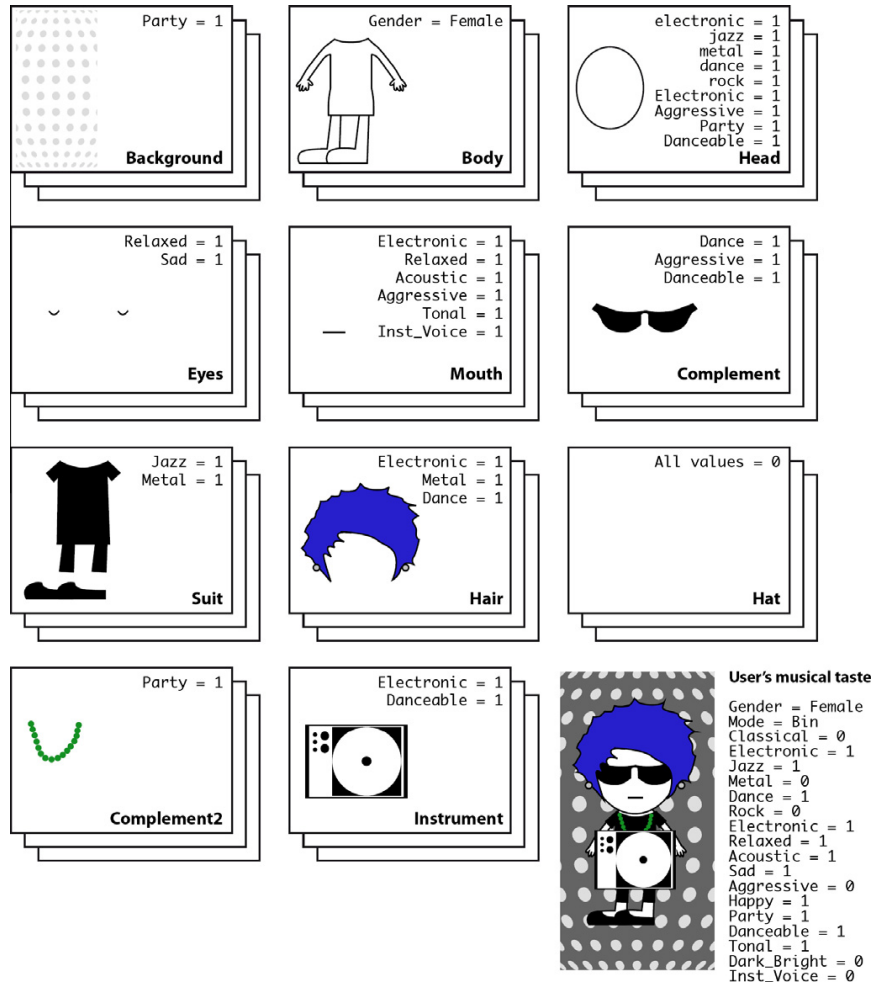


Fig. 6. Example of the visualization approach. It can be seen how the descriptor values influence the selection of the different graphic elements used to construct the avatar. The values inside the graphic element boxes represent all possible descriptor values that can generate the presented element.

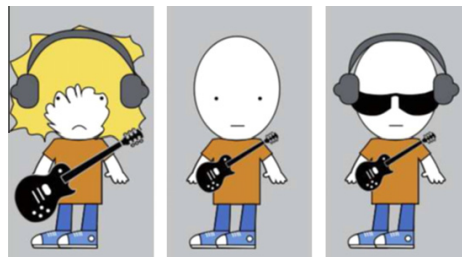


Fig. 7. Sample Musical Avatars generated by the three summarization methods (i.e., from left to right, binary, ternary, and continuous) for the same underlying user model. Notice the differences in guitar and headphones sizes among the generated avatars.

each descriptor (e.g., 0 meaning “I do not like classical music at all” up to 1 meaning “I like classical music a lot” in the case of the “classical” descriptor). For the second task, we first showed 20 randomly generated examples of the Musical Avatars in order to introduce their visual nature. We then presented to each participant six avatars: namely, the three images generated

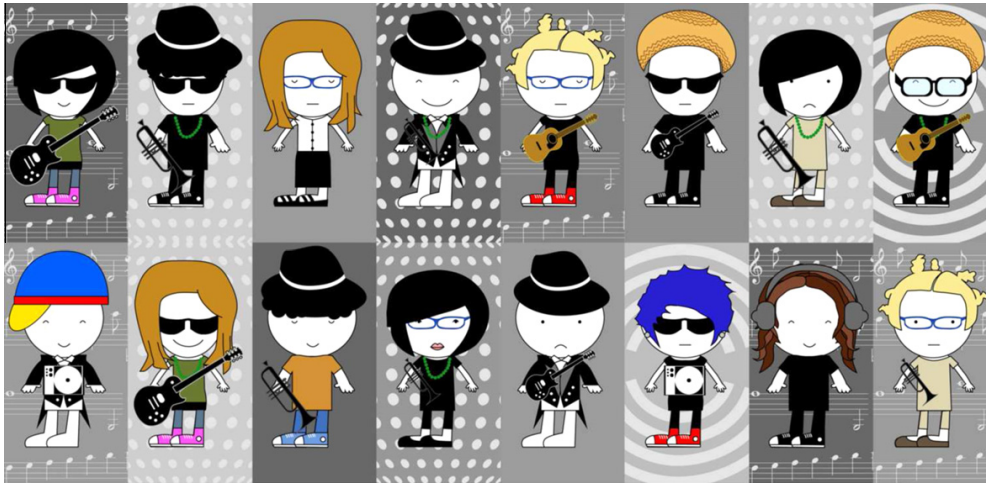


Fig. 8. A random sample of Musical Avatars.

Table 7

Mean ranks and standard deviations for the different visualization methods obtained in the user evaluation. The random column corresponds to the average values of the individual random results (see text for details).

	Continuous	Binary	Ternary	Random	Neutral
$\mu$	1.73	2.27	2.91	4.28	5.18
$\sigma$	0.79	1.49	1.45	1.16	0.98

from her/his own preference set, two randomly generated avatars, and one neutral avatar. We asked the participants to rank these images assigning the image that best express their musical preferences to the first position in the rank (i.e., rank = 1). Finally, we asked for a written feedback regarding the images, the evaluation procedure, or any other comments.<sup>20</sup>

### 6.3.2. Results

From the obtained data we first analyzed the provided rankings to estimate the accuracy of the visualization methods examined in the questionnaire. To this end, we computed the mean rank for each method. The resulting means and standard deviations are reported in Table 7. We tested the effect of the method on the ratings obtained from the subjects using a within-subjects ANOVA. The effect of the visualization method was found to be significant (Wilks Lambda = 0.032,  $F(4,7) = 52,794$ ,  $p < 0.001$ ). Pairwise comparisons (a least significant differences t-test with Bonferroni correction, which conservatively adjusts the observed significance level based on the fact that multiple comparisons are made) revealed significant differences between two groups of avatars: on one side, the random and the neutral avatars (getting ratings that cannot be considered different from each other) and, on the other side, the *binary*, *ternary*, and *continuous* avatars (which get ratings that are statistically different from the random and the neutral ones, but without any significant difference between the three). The differences between those two groups of avatars are clearly significant ( $p < 0.005$ ) except for the differences between random and *ternary*, and between *binary* and neutral, which are only marginally significant ( $p \leq 0.01$ ).

We then introduced a dissimilarity measure to assess the significance of the summarized description of musical preferences. In particular, we estimated how the computed representation performs against a randomly generated baseline. Therefore, we first computed the Euclidean distance between the obtained descriptor vector representing the user profile (standardized and range-normalized) and the vector containing the participants' self-assessments provided in the first task of the evaluation. We then generated a baseline by averaging the Euclidean distances between the self-assessments and 10 randomly generated vectors. Finally, a t-test between the algorithm's output ( $\mu = 0.99$ ,  $\sigma = 0.32$ ) and the baseline ( $\mu = 1.59$ ,  $\sigma = 0.25$ ) showed a significant difference in the sample's means ( $t(11) = -5.11$ ,  $p < 0.001$ ).

From the obtained results, we first observe that the generated description based on audio content analysis shows significant differences when compared to a random assignment. The mean distance to the user-provided values is remarkably smaller for the generated data than for the random baseline; i.e., the provided representations reasonably approximate the users' self-assessments in terms of similarity. Furthermore, Table 7 clearly shows a user preference for all three proposed

<sup>20</sup> A screenshot of the evaluation and more Musical Avatars are available online <http://mtg.upf.edu/project/musicalavatar>.

quantization methods over the randomly generated and the neutral *Musical Avatars*. In particular, the *continuous* summarization method has been found top-ranked, followed by the *binary* and *ternary* quantization methods. This ranking, given the ANOVA results, should be taken just as approximative until a larger sample of user evaluations is available. Specifically, the conducted ANOVA did not reveal a clear particular preference for any of the three considered methods.

Evaluation of the participants' comments can be summarized as follows. First, we can observe a general tendency towards an agreement on the representativeness of the *Musical Avatar*. As expected, some subjects reported missing categories to fully describe their musical preferences (e.g., country music, musical instruments). This suggests that the provided semantic descriptors seem to grasp the essence of the user's musical preference, but fail to describe subtle nuances in detail. This could be explained by the fact that we use a reduced set of semantic descriptors in our prototype (17 descriptors out of the 62 initially extracted for the proposed user model). Finally, some participants could not decode the meaningfulness of some visual features (e.g., glasses, head shape). This information will be considered in our future work for refining the mapping strategy. According to the obtained results, we observed participants' preference for all three summarization methods based on the proposed user model over the baselines. In general, we conclude that the *Musical Avatar* provides a reliable, albeit coarse, visual representation of the user's musical preferences.

## 7. General discussion and possible enhancements for recommender systems

Let us shortly recapitulate the major contents of the current work. We proposed a novel technique for semantic preference elicitation suitable for various applications within music recommender systems and information filtering systems in general. Two such applications – music recommendation and musical preference visualization, were presented and evaluated from a user modeling point-of-view. Moreover, the proposed user modeling approach and the considered applications can be used as basic tools for human computer interaction to enrich the experience with music recommender systems. A number of innovative personalized interfaces for understanding, discovering, and manipulating music recommendations can be built on top of our developed methodologies.

Considering the limitations of our study, we would like to note that we employed a small sample of subjects (12 music enthusiasts) that might not represent the general population. We nevertheless observed statistical significant differences which, in this context, mean that the detected trends are strong enough to override the individual differences or potentially large variability that might be observed in small-size samples of listeners. We also believe that users of music recommender systems, at least to date, are mainly music enthusiasts, and hence we have properly and sufficiently sampled that population. More importantly, to the best of our knowledge, the few existing research studies on music recommendation involving evaluations with real participants are significantly limited in the tradeoff between the number of participants (Hoashi et al., 2003) and the number of evaluated tracks per approach by a particular user Barrington et al. (2009) and Lu and Tseng (2009). Furthermore, no studies on human evaluation of visualization approaches considering musical preferences are known to the authors.

In what follows we comment on the implications of the presented approaches for the user's interaction as well as future implementations of "final systems", which unite both applications, recommendation and visualization, into a single, interactive music recommender interface. For both considered applications a user-defined preference set served as a starting point. This preference set is automatically converted into a fully semantic user model. The preference set offers a compact description of presumably multi-faceted preferences explicitly in terms of multiple music tracks. Therefore it is not limited to a single seed item or semantic term to draw recommendations from. In contrast, the preference set manually provided by the user assures a noise-free representation of the user's preference with a maximum possible coverage. Moreover, the chosen preference elicitation technique – namely inferring the dimensions of the user's preferences in a fully audio content-based manner – provides the system with the flexibility to overcome the so-called cold-start problem, which audio content-unaware systems are typically faced with (see Section 1). It also guarantees recommendations of non-popular items, which may be preferred by specialized or long-tail seeking listeners. Finally, having semantic descriptions for both the user and the recommended items allows to automatically generate justifications for the recommended music (e.g., "This track was recommended because you like jazz music with acoustic instrumentation and relaxed mood"), which is a highly desirable feature for a recommendation system (Tintarev & Masthoff, 2007).

The mapping of the semantic dimensions to visual features, resulting in the *Musical Avatar*, enables an intuitive, yet still arbitrary, depiction of musical preferences. This by itself enriches and facilitates the user's interaction process, an appealing feature for any recommender system. Furthermore, allowing the user to interact and manipulate graphical representations offers a straightforward path towards user adaptive models. One possible extension here is the filtering of music recommendations according to the presence or absence of certain visual features of the *Musical Avatar*. This allows users to actively control the output of the music recommender by selecting certain visual attributes which are connected to acoustic properties via the mapping described in Section 6. Also, the iconic *Musical Avatar* may serve as a badge, reflecting a quick statement of one's musical preferences, with possible applications in online social interaction. Moreover, users can share preferences related to the generated avatars or group together according to similar musical preferences represented by the underlying user models.

Both aforementioned applications can be easily united into a single interactive recommender system. In addition to the already discussed music recommendation and static preference visualization, the concepts introduced in the present work



can be extended to reflect time-varying preferences. For example, an underlying user model can be computed considering different time periods (e.g., yesterday, last week, last month). Also, tracking preferences over time enables the generation of “preference time-lines”, where *Musical Avatars* morph from one period to the next, while users can ask for recommendations from different periods of their musical preferences.

Moreover, in the visualization application, exploiting multiple instances of preference sets can alleviate the limitations introduced by a single preference set. Multiple graphical instances can be used to visually describe different subsets of a music collection, thus serving as high-level tools for media organization and browsing. Hence, recommendations can be directed by those avatars, introducing one additional semantic visual layer in the recommendation process. Using multiple representations can help to better visually depict preferences of certain users, where a single avatar is not sufficient for describing all facets of their musical preferences. Moreover, users may want to generate context dependent avatars, which can be used for both re-playing preference items or listening to recommendations depending on the context at hand (e.g., one may use his avatar for happy music at a party or listen to recommendations from the “car” avatar while driving).

Finally, alternative methods for gathering the preference set can be employed. Since selecting representative music tracks may be a boring and exhausting task for certain users, data-driven approaches can be applied. Audio content-based methods may be used to infer preference items from the user’s personal collection by, for instance, clustering the collection according to certain musical facets to find central elements within each cluster (i.e., centroids). Additionally, listening statistics or personal ratings of particular items can be used to infer musical preferences without actually processing a full music collection.<sup>21</sup> Nevertheless, such an implicit inference of a preference set can lead to noisy representations or to the lack of coverage of all possible facets of the user’s musical preferences (see also Section 1).

## 8. Conclusions

In the present work we considered audio content-based user modeling approaches suitable for music recommendation and visualization of musical preferences. We proposed a novel technique for preference elicitation, which operates on an explicitly given set of music tracks defined by a user as evidence of her/his musical preferences and builds a user model by automatically extracting a semantic description of the audio content for each track in the set. To demonstrate the effectiveness of the proposed technique we considered (and evaluated) two applications: music recommendation and visualization of musical preferences. The results obtained from the subjective evaluations, conducted on 12 subjects, are promising.

In the case of music recommendation, we demonstrated that the approaches based on the proposed semantic model, inferred from low-level timbral, temporal, and tonal features, outperform state-of-the-art audio content-based algorithms exploiting only low-level timbral features. Although these approaches perform worse than the considered commercial black-box system, which exploit collaborative filtering, the difference in performance is greatly diminished when using the semantic descriptors computed in our model. It is important to notice that one of the main advantages of our model is the fact that it does not suffer from the long-tail and cold-start problems, which are inherent to collaborative filtering approaches.

In the case of musical preferences visualization, we presented an approach to automatically create a visual avatar of the user, capturing their musical preferences, from the proposed user model. To the best of our knowledge, such a task has not been previously explored in the literature, and we have developed an appropriate procedure and an evaluation methodology. The subjective evaluation showed that the provided visualization is able to reliably depict musical preferences, albeit in a coarse way.

In addition to the demonstrated applications, we also described a number of possible enhancements of music recommender systems based on the proposed user model. Specifically, we discuss justification of recommendations, interactive interfaces based on visual clues, playlist description and visualization, tracking the evolution of a user’s musical preferences, and social applications.

As future work, we plan to focus our research on performance improvements, enriching the current model with more semantic descriptors (e.g., instrument information), and improving the accuracy of the underlying classifiers. We also plan to expand the present prototypical study and conduct a large scale Web-based user evaluation in order to better assess the representativeness of the obtained user models for their further refinement. In particular, as the proposed technique requires some effort from the user to gather preference examples, a comparison with implicit methods to obtain information about preferences would be of interest.

## Acknowledgements

The authors thank all participants involved in the evaluation and Justin Salamon for proofreading. This research has been partially funded by the FI Grant of Generalitat de Catalunya (AGAUR) and the Buscamedia (CEN-20091026), Classical Planet (TSI-070100-2009-407, MITYC), DRIMS (TIN2009-14247-C02-01, MICINN), and MIRES (EC-FP7 ICT-2011.1.5 Networked Media and Search Systems, grant agreement No. 287711) Projects.

<sup>21</sup> A demo of such a music recommender/visualization system working on the proposed principles, but taking listening statistics instead of explicitly given preference set, is available online <http://mtg.upf.edu/project/musicalavatar>.

## References

- Abdullah, M. B. (1990). On a robust correlation coefficient. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 39(4), 455–460.
- Ahmed, N., de Aguiar, E., Theobalt, C., Magnor, M., & Seidel, H. (2005). Automatic generation of personalized human avatars from multi-view video. In *ACM Symp. on virtual reality software and technology (VRST'05)* (pp. 257–260).
- Amatriain, X., Pujol, J., & Oliver, N. (2009). I like it... i like it not: Evaluating user ratings noise in recommender systems. *User Modeling, Adaptation, and Personalization*, 5535/2009, 247–258.
- Aucouturier, J. J. (2009). Sounds like teen spirit: Computational insights into the grounding of everyday musical terms. In J. Minett & W. Wang (Eds.), *Language, evolution and the brain. Frontiers in linguistics* (pp. 35–64). Taipei: Academia Sinica Press.
- Baltrunas, L., & Amatriain, X. (2009). Towards time-dependant recommendation based on implicit feedback. In *Workshop on context-aware recommender systems (CARS'09)*.
- Barrington, L., Oda, R., & Lanckriet, G. (2009). Smarter than genius? Human evaluation of music recommender systems. In *Int. society for music information retrieval conf. (ISMIR'09)* (pp. 357–362).
- Barrington, L., Turnbull, D., Torres, D., & Lanckriet, G. (2007). Semantic similarity for music retrieval. In *Music information retrieval evaluation exchange (MIREX'07)*. <[http://www.music-ir.org/mirex/abstracts/2007/AS\\_barrington.pdf](http://www.music-ir.org/mirex/abstracts/2007/AS_barrington.pdf)>.
- Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.
- Bogdanov, D., Haro, M., Fuhrmann, F., Gómez, E., & Herrera, P. (2010). Content-based music recommendation based on user preference examples. In *ACM conf. on recommender systems. Workshop on music recommendation and discovery (Womrad 2010)*.
- Bogdanov, D., Serrà, J., Wack, N., & Herrera, P. (2009). From low-level to high-level: Comparative study of music similarity measures. In *IEEE int. symp. on multimedia (ISM'09). Int. workshop on advances in music information research (AdMIR'09)* (pp. 453–458).
- Bogdanov, D., Serrà, J., Wack, N., Herrera, P., & Serra, X. (2011). Unifying low-level and high-level music similarity measures. *IEEE Transactions on Multimedia*, 13(4), 687–701.
- Brossier, P.M. (2007). *Automatic annotation of musical audio for interactive applications*. Ph.D. thesis, QMUL, London, UK.
- Cano, P., Gómez, E., Gouyon, F., Herrera, P., Koppenberger, M., Ong, B. et al. (2006). *ISMIR 2004 audio description contest*. Tech. rep. <<http://mtg.upf.edu/node/461>>.
- Casey, M. A., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., & Slaney, M. (2008). Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE*, 96(4), 668–696.
- Celma, O. (2008). *Music recommendation and discovery in the long tail*. Ph.D. thesis, UPF, Barcelona, Spain.
- Celma, O., & Herrera, P. (2008). A new approach to evaluating novel recommendations. In *ACM conf. on recommender systems (RecSys'08)* (pp. 179–186).
- Celma, O., Herrera, P., & Serra, X. (2006). Bridging the music semantic gap. In *ESWC 2006 workshop on mastering the gap: From information extraction to semantic representation*. <<http://mtg.upf.edu/node/874>>.
- Celma, O., & Serra, X. (2008). FOAFing the music: Bridging the semantic gap in music recommendation. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(4), 250–256.
- Chang, C., & Lin, C. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3), 27:1–27:27.
- Cosley, D., Lam, S. K., Albert, I., Konstan, J. A., & Riedl, J. (2003). Is seeing believing?: How recommender system interfaces affect users' opinions. In *Conf. on human factors in computing systems (CHI'03)* (pp. 585–592).
- Cramer, H., Evers, V., Ramlal, S., Someren, M., Rutledge, L., Stash, N., et al (2008). The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction*, 18(5), 455–496.
- Downie, J., Ehmann, A., Bay, M., & Jones, M. (2010). The music information retrieval evaluation eXchange: some observations and insights. In *Advances in music information retrieval* (pp. 93–115).
- Firan, C. S., Nejd, W., & Paiu, R. (2007). The benefit of using tag-based profiles. In *Latin American web conf.* (pp. 32–41).
- Grimaldi, M., & Cunningham, P. (2004). Experimenting with music taste prediction by user profiling. In *ACM SIGMM int. workshop on multimedia information retrieval (MIR'04)* (pp. 173–180).
- Grimmett, G., & Stirzaker, D. (2001). *Probability and random processes* (3rd ed.). Oxford University Press.
- Gómez, E. (2006). *Tonal description of music audio signals*. Ph.D. thesis, UPF, Barcelona, Spain.
- Gómez, E., & Herrera, P. (2008). Comparative analysis of music recordings from western and Non-Western traditions by automatic tonal feature extraction. *Empirical Musicology Review*, 3(3), 140–156.
- Hall, M. A. (2000). Correlation-based feature selection for discrete and numeric class machine learning. In *Int. conf. on machine learning* (pp. 359–366).
- Hanani, U., Shapira, B., & Shoval, P. (2001). Information filtering: Overview of issues, research and systems. *User Modeling and User-Adapted Interaction*, 11(3), 203–259.
- Haro, M., Xambó, A., Fuhrmann, F., Bogdanov, D., Gómez, E., & Herrera, P. (2010). The Musical Avatar – a visualization of musical preferences by means of audio content description. In *Audio Mostly (AM '10)*.
- Hoashi, K., Matsumoto, K., & Inoue, N. (2003). Personalization of user profiles for content-based music retrieval based on relevance feedback. In *ACM int. conf. on multimedia (MULTIMEDIA'03)* (pp. 110–119).
- Holm, J., Aaltonen, A., & Siirtola, H. (2009). Associating colours with musical genres. *Journal of New Music Research*, 38(1), 87–100.
- Homburg, H., Mierswa, I., Möller, B., Morik, K., & Wurst, M. (2005). A benchmark dataset for audio classification and clustering. In *Int. conf. on music information retrieval (ISMIR'05)* (pp. 528–531).
- Jawaheer, G., Szomszor, M., & Kostkova, P. (2010). Comparison of implicit and explicit feedback from an online music recommendation service. In *Int. Workshop on information heterogeneity and fusion in recommender systems (HetRec'10). HetRec'10* (pp. 47–51). New York, NY, USA: ACM. ACM ID: 1869453.
- Jones, N., & Pu, P. (2007). User technology adoption issues in recommender systems. In *Networking and electronic commerce research conf.*
- Laurier, C., Meyers, O., Serrà, J., Blech, M., & Herrera, P. (2009). Music mood annotator design and integration. In *Int. workshop on content-based multimedia indexing (CBMI'2009)*.
- Levy, M., & Bosteels, K. (2010). Music recommendation and the long tail. In *ACM conf. on recommender systems. workshop on music recommendation and discovery (Womrad 2010)*.
- Li, Q., Myaeng, S., Guan, D., & Kim, B. (2005). A probabilistic model for music recommendation considering audio features. In *Information retrieval technology* (pp. 72–83).
- Li, Q., Myaeng, S. H., & Kim, B. M. (2007). A probabilistic music recommender considering user opinions and audio features. *Information Processing and Management*, 43(2), 473–487.
- Logan, B. (2004). Music recommendation from song sets. In *Int. conf. on music information retrieval (ISMIR'04)* (pp. 425–428).
- Lu, C., & Tseng, V. S. (2009). A novel method for personalized music recommendation. *Expert Systems with Applications*, 36(6), 10035–10044.
- MacDonald, R. A. R., Hargreaves, D. J., & Miell, D. (2002). *Musical identities*. Oxford University Press.
- Maffesoli, M. (1996). *The time of the tribes: the decline of individualism in mass society*. SAGE.
- Maltz, D., & Ehrlich, K. (1995). Pointing the way: active collaborative filtering. In *SIGCHI conf. on human factors in computing systems (CHI'95)* (pp. 202–209).
- Mandel, M. I., & Ellis, D. P. (2005). Song-level features and support vector machines for music classification. In *Int. conf. on music information retrieval (ISMIR'05)* (pp. 594–599).
- McCloud, S. (2009). *Understanding comics: The invisible art* (36th ed.). HarperPerennial.
- Nanopoulos, A., Rafailidis, D., Ruxanda, M., & Manolopoulos, Y. (2009). Music search engines: Specifications and challenges. *Information Processing and Management*, 45(3), 392–396.

- Pampalk, E. (2006). *Computational models of music similarity and their application in music information retrieval*. Ph.D. thesis, Vienna University of Technology.
- Peeters, G. (2004). *A large set of audio features for sound description (similarity and classification) in the CUIDADO project*. CUIDADO Project Report.
- Petajan, E. (2005). Face and body animation coding applied to HCL. In *Real-time vision for human-computer interaction*. US: Springer.
- Platt, J. C. (2000). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In P. J. Bartlett, B. Schölkopf, D. Schuurmans, & A. J. Smola (Eds.), *Advances in large margin classifiers* (pp. 61–74). Cambridge, MA: MIT Press.
- Pohle, T., Schnitzer, D., Schedl, M., Knees, P., & Widmer, G. (2009). On rhythm and general music similarity. In *Int. society for music information retrieval conf. (ISMIR'09)* (pp. 525–530).
- Reas, C., & Fry, B. (2007). *Processing: A programming handbook for visual designers and artists*. MIT Press.
- Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11), 613–620.
- Sarwar, B., Karypis, G., Konstan, J., & Reidl, J. (2001). Item-based collaborative filtering recommendation algorithms. In *Int. conf. on World Wide Web (WWW'01)* (pp. 285–295).
- Sauer, D., & Yang, Y. (2009). Music-driven character animation. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 5(4), 1–16.
- Schedl, M., Widmer, G., Knees, P., & Pohle, T. (2011). A music information system automatically generated via web content mining techniques. *Information Processing and Management*, 47(3), 426–439.
- Shardanand, U., & Maes, P. (1995). Social information filtering: algorithms for automating “word of mouth”. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 210–217).
- Sotiropoulos, D. N., Lampropoulos, A. S., & Tsihrintzis, G. A. (2007). MUSIPER: A system for modeling music similarity perception based on objective feature subset selection. *User Modeling and User-Adapted Interaction*, 18(4), 315–348.
- Streich, S. (2007). *Music complexity: a multi-faceted description of audio content*. Ph.D. thesis, UPF, Barcelona, Spain.
- Su, J. H., Yeh, H. H., & Tseng, V. S. (2010). A novel music recommender by discovering preferable perceptual-patterns from music pieces. In *ACM symp. on applied computing (SAC'10)* (pp. 1924–1928).
- Tintarev, N., & Masthoff, J. (2007). Effective explanations of recommendations: user-centered design. In *Proceedings of the 2007 ACM conference on recommender systems* (pp. 153–156).
- Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5), 293–302.
- Vignoli, F., & Pauws, S. (2005). A music retrieval system based on user-driven similarity and its evaluation. In *Int. conf. on music information retrieval (ISMIR'05)* (pp. 272–279).
- Voong, M., & Beale, R. (2007). Music organisation using colour synaesthesia. In *CHI'07 extended abstracts on Human Factors in Computing Systems* (pp. 1869–1874).
- West, K., & Lamere, P. (2007). A model-based approach to constructing music similarity functions. *EURASIP Journal on Advances in Signal Processing*, 2007, 149.
- Yoshii, K., Goto, M., Komatani, K., Ogata, T., & Okuno, H. G. (2006). Hybrid collaborative and content-based music recommendation using probabilistic model with latent user preferences. In *Int. conf. on music information retrieval (ISMIR'06)*.

**Herrera, P.**, Resa Z., & Sordo M. (2010). Rocking around the clock eight days a week: an exploration of temporal patterns of music listening. 1st Workshop on Music Recommendation and Discovery (WOMRAD), ACM RecSys, 2010, Barcelona.

[http://mtg.upf.edu/files/publications/womrad2010\\_submission\\_16.pdf](http://mtg.upf.edu/files/publications/womrad2010_submission_16.pdf)

# Rocking around the clock eight days a week: an exploration of temporal patterns of music listening

Perfecto Herrera

Zuriñe Resa

Mohamed Sordo

Music Technology Group  
Department of Technology  
Universitat Pompeu Fabra

perfecto.herrera@upf.edu

zuri\_resa@hotmail.com

mohamed.sordo@upf.edu

## ABSTRACT

Music listening patterns can be influenced by contextual factors such as the activity a listener is involved in, the place one is located or physiological constants. As a consequence, musical listening choices might show some recurrent temporal patterns. Here we address the hypothesis that for some listeners, the selection of artists and genres could show a preference for certain moments of the day or for certain days of the week. With the help of circular statistics we analyze playcounts from Last.fm and detect the existence of that kind of patterns. Once temporal preference is modeled for each listener, we test the robustness of that using the listener's playcount from a posterior temporal period. We show that for certain users, artists and genres, temporal patterns of listening can be used to predict music listening selections with above-chance accuracy. This finding could be exploited in music recommendation and playlist generation in order to provide user-specific music suggestions at the "right" moment.

## Categories and Subject Descriptors

H.5.5 Sound and Music Computing – methodologies and techniques, modeling.

## General Terms

Measurement, Experimentation, Human Factors.

## Keywords

Music context analysis, Playlist generation, User modeling, Music metadata, Temporal patterns, Music preference.

## 1. INTRODUCTION

Among the requirements of good music recommenders we can point to, not only delivering the right music but, delivering it at the right moment. This amounts to consider the context of listening as a relevant variable in any user model for music recommendation. As existing technologies also make it possible to track the listening activity every time and everywhere it is happening, it seems pertinent to ask ourselves how this tracking can be converted into usable knowledge for our recommendation

*WOMRAD 2010* Workshop on Music Recommendation and Discovery, collocated with ACM RecSys 2010 (Barcelona, SPAIN).

Copyright ©. This is an open-access article distributed under the terms of the Creative Commons Attribution License 3.0 Unported, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

systems. Music listening decisions might seem expressions of free will but they are in fact influenced by interlinked social, environmental, cognitive and biological factors [21][22].

Chronobiology is the discipline that deals with time and rhythm in living organisms. The influence of circadian rhythms (those showing a repetition pattern every 24 hours approximately, usually linked to the day-night alternation), but also of ultradian rhythms (those recurring in a temporal lag larger than one day like the alternation of work and leisure or the seasons), has been demonstrated on different levels of organization of many living creatures, and preserving some biological cycles is critical to keep an optimum health [18]. The observation that human behavior is modulated by rhythms of hormonal releases, exposure to light, weather conditions, moods, and also by the activity we are engaged into [12][3] paves the way to our main hypothesis: there are music listening decisions that reflect the influence of those rhythms and therefore show temporal patterns of occurrence. The connection would be possible because of the existing links between music and mood on one side, and between music and activity on the other side. In both cases, music has functional values either as mood regulator [23] or as an activity regulator [13]. Therefore, as mood and activity are subject to rhythmic patterns and cycles, music selection expressed in playlists could somehow reflect that kind of patterning [26][23]. More specifically, in this paper we inquire on the possibility of detecting that, for a specific user, certain artists or musical genres are preferentially listened to at certain periods of the day or on specific days of the week. The practical side of any finding on this track would be the exploitation of this knowledge for a better contextualized music recommendation. Our research is aligned with a generic trend on detecting hidden patterns of human behavior at the individual level thanks, mainly, to the spread of portable communication and geolocation technologies [4][20].

## 2. RELATED RESEARCH

While recommendations based on content analysis or on collaborative filtering may achieve a certain degree of personalization, they do miss the fact that the users interact with the systems in a particular context [19]. Furthermore, several studies have shown that a change in contextual variables induces changes in user's behaviors and, in fact, when applying contextual modelling of the users (i.e., considering the time of the day, the performed activity, or the lighting conditions), the performance of recommendation systems improves both in terms of predictive accuracy and true positive ratings [8][25]. Although context-based music recommenders were available since 2003 [1], time information is a recently-added contextual feature [7][17].

A generic approach to the characterization of temporal trends in everyday behavior has been presented in [10], where the concept of “eigenbehavior” is introduced. Eigenbehaviors are characteristic behaviors (such as leaving early home, going to work, breaking for lunch and returning home in the evening) computed from the principal components of any individual’s behavioral data. It is an open research issue if Eigenbehaviors could provide a suitable framework for analyzing music listening patterns. A model tracking the time-changing behavior of users and also of recommendable items throughout the life span of the data was developed for the Netflix movie collection [14]. This allowed the author to detect concept drifts and the temporal evolution of preferences, and to improve the recommendation over a long time span.

Although research on behavioral rhythms has a long and solid tradition, we are not aware of many studies about their influence on music listening activities. The exception is a recent paper [2] where users’ *micro-profiles* were built according to predefined non-overlapping temporal partitions of the day (e.g., “morning time slot”). The goal of the authors was to build a time-aware music recommender and their evaluation of the computed micro-profiles showed their potential to increase the quality of recommendations based on collaborative filtering. Most of that reported work was, though, on finding optimal temporal partitions. As we will see, there are other feasible, maybe complementary, options that keep the temporal dimension as a continuous and circular one by taking advantage of circular statistics. Developed forty years ago and largely used in biological and physical sciences, circular statistics has also been exploited in personality research for studying temporal patterns of mood [15][16]. To our knowledge, it is the first time they are used in the analysis of music-related behavior, though applications to music have been previously reported [5][9].

### 3. METHODOLOGY

#### 3.1 Data Collection

Getting access to yearly logs of the musical choices made by a large amount of listeners is not an easy task. Many music playing programs store individual users’ records of that, but they are not publicly accessible. As a workable solution, we have taken advantage of Last.fm API, which makes possible to get the playcounts and related metadata of their users. As raw data we have started with the full listening history of 992 unique users, expressed as 19,150,868 text lines and spanning variable length listening histories from 2005 to 2009. The data contained a user identifier, a timestamp, Musicbrainz identifiers for the artist and track, and a text name for the listened track.

The artist genre information was gathered from Last.fm using the Last.fm API method *track.getTopTags()*, which returns a list of tags and their corresponding weight<sup>1</sup>. This list of tags, however, may relate to different aspects of music (e.g. genre, mood, instrumentation, decades...). Since in our case we need a single genre per track, we first clean tags in order to remove special characters or any other undesirable characters, such as spaces, hyphens, underscores, etc. Then irrelevant tags (i.e., those having

a low weight) are removed and the remaining ones are matched against a predefined list of 272 unique musical genres/styles gathered from Wikipedia and Wordnet. From the genre tags we obtained for each song, we select the one with the highest weight. If there are several tags with the highest weight, we select the one with the least popularity (popularity is computed as the number of occurrences of a specific genre in our data-set).

#### 3.2 Data cleaning

Data coming from Last.fm.com contain playcounts that cannot be attributable to specific listening decisions on the side of users. If they select radio-stations based on other users, on tags or on similar artists there are chances that songs, artists and genres will not recur in a specific user’s profile. In general, even in the case of having data coming from personal players obeying solely to the user’s will, we should discard (i) users that do not provide enough data to be processed, and (ii) artists and genres that only appear occasionally. We prefer to sacrifice a big amount of raw data provided those we keep help to identify a few of clearly recurring patterns, even if it is only for a few users, artists or genres.

In order to achieve the above-mentioned cleaning goals we first compute, for each user, the average frequency of each artist/genre in his/her playlist. Then, for each user’s dataset, we filter out all those artists/genres for which the playlist length is below the user’s overall average playlist length. Finally, in order to get rid of low-frequency playing users, we compute the median value of the number of artists/genres left after the last filtering step, which we will name as “valid” artists/genres. Those users whose number of “valid” artists/genres is below the median percentage value are discarded.

#### 3.3 Prediction and Validation Data Sets

Once we get rid of all the suspected noise, we split our dataset in two groups. One will be used to generate the temporal predictions while the other one will be used to test them. The test set contains all the data in the last year of listening for a given subject. The prediction-generation set contains the data coming from two years of listening previous to the year used in the test set.

#### 3.4 Circular Statistics

Circular statistics are aimed to analyze data on circles where angles have a meaning, which is the case when dealing with daily or weekly cycles. In fact, circular statistics is an alternative to common methods or procedures for identifying cyclic variations or patterns, which include spectral analysis of time-series data or time-domain based strategies [15]. Although these approaches are frequently used, their prerequisites (e.g., interval scaling, regularly spaced data, Gaussianity) are seldom met and, as we mentioned above, these techniques have rarely been used to analyze music-related data and therefore we wanted to explore its potential.

Under the circular statistics framework, variables or data considered to be cyclic in nature are meant to have a period of measurement that is rotationally invariant. In our case this period is referred to the daily hours and the days of the week. Therefore, taking into account the rotationally invariant period of analysis this would be reflected as daily hours that range from 0 to 24, where 24 is considered to be the same as 0. Regarding to the weekly rhythm, Monday at 0h would be considered to be the same as Sunday at 24h.

---

<sup>1</sup> Last.fm relevance weight of tag  $t$  to artist  $a$ , ranging from 0 to 100.

The first step in circular analysis is converting raw data to a common angular scale. We chose the angular scale in radians, and thus we apply the following conversion to our dataset:

$$\alpha = \frac{2\pi x}{k}$$

where  $x$  represents raw data in the original scale,  $\alpha$  is its angular direction (in radians) and  $k$  is the total number of steps on the scale where  $x$  is measured. In fact, we denote  $\alpha$  as a vector of  $N$  directional observations  $\alpha_i$  ( $i$  ranging from 1 to  $N$ ). For the daily hour case,  $x$  would have values between 0 and 24, and  $k = 24$ . Alternatively, for the weekday analysis,  $x$  would have a scale from 0 (Monday) to 6 (Sunday) and thus,  $k = 6$ . As noted, the effect of this conversion can be easily transformed back to the original scale. Once we have converted our data to angular scale, we compute the *mean direction* (a central tendency measure) by transforming raw data into unit vectors in the two-dimensional plane by

$$r_i = \begin{pmatrix} \cos \alpha_i \\ \sin \alpha_i \end{pmatrix}$$

After this transformation, vectors  $r_i$  are vector-averaged by

$$\bar{r} = \frac{1}{N} \sum_i r_i$$

The quantity  $\bar{r}$  is the *mean resultant vector* associated to the mean direction, and its length  $\bar{R}$  describes the spread of the data around the circle. For events occurring uniformly in time  $\bar{R}$  values approach 0 (uniform circular distribution) whereas events concentrated around the mean direction yield values close to 1 (see figure 1 for an example). A null hypothesis (e.g., uniformity) about the distribution of data can be assessed using Rayleigh's [11] or Omnibus (Hodges-Ajne) tests [27], the latter working well for many distribution shapes. Once we have detected significantly modally distributed data by means of both tests, we verify that it wasn't completely pointing to a single day or hour. All the circular statistics analyses presented here have been performed with the CircStat toolbox for Matlab [6].

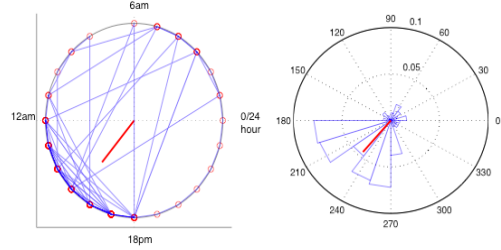
## 4. RESULTS

### 4.1 Data cleaning

As a consequence of the cleaning process, our working dataset now contains data from 466 valid users. The cleaning process has kept 62% of their total playcounts, which corresponds to 4.5% of the initial amount of artists. This dramatic reduction of the artists should not be surprising as many listening records show a "long-tail" distribution, with just a few of frequently played artists, and many of them seldom played. On the other hand, when focusing on musical genre listening, the working dataset includes 515 users, from which 78% of their playcounts has been kept. These playcounts comprise 8.6% of the total number of genres. Again, a long-tail distribution of the amount of listened genres is observed.

### 4.2 Temporal Patterns of Artist Selection

Once we have cleaned our dataset, we compute the mean circular direction and the mean resultant vector length for each artist and user. Therefore, these values can be considered as a description of the listening tendencies for each artist by each user. Both parameters were calculated for the daily and for the weekly data.



**Figure 1. Circular representation of a specific user listening behavior for a specific artist along 24 hours. The left side diagram shows the daily distribution of listening, and the right one the circular histogram. The red line represents the mean vector direction and length in both cases.**

In order to assess the relevance of these listening trends, we tested that the distribution of playcounts was different from uniform, and that it was modally distributed (i.e. showing a tendency around an hour or around a day of the week) and discarded those that were not fulfilling these requirements (a null hypothesis rejection probability  $p < 0.05$  was set for the tests).

In the hour prediction problem, for each listener's clean dataset almost 93% ( $\sigma=13$ ) of the artists passed on average the uniformity test (i.e., listening to them is meant to be concentrated around a specific hour). However, considering the raw dataset, only a per-user average of 7% ( $\sigma=3.2$ ) of the artists show a listening hour tendency. For the weekly approach, the per-user average in the clean dataset is 99.8% ( $\sigma=0.8$ ), indicating that there are some artists showing a clear tendency towards a preferred listening day. Considering the original raw dataset, they correspond to a 7.5% ( $\sigma=3.2$ ) of all the played artists.

Data from 466 users, including 7820 different songs and a grand total of 23669 playcounts were used in the validation of the temporal listening patterns of artists. For each user and artist we computed a "hit" if the absolute difference between the playing day in the prediction and test conditions, expressed as a circular mean value in radians, was less than 0.45 (the equivalent to a half-a-day error). For the time of the day a half-an-hour error was accepted, corresponding to a difference between the predicted and the observed time of less than 0.13 radians.

When predicting the day of listening, an overall 32.4% of hits was found for the songs in the test collection, which exceeds by far the chance expectations ( $1/7=14.28\%$ ). As the final goal of the model is providing user-specific contextual recommendation, an additional per-user analysis yielded 34.5% of hits ( $\sigma=17.8$ ). Identical data treatment was done with the time of the day yielding an overall 17.1% of hits (chance expectation baseline:  $1/24=4.1\%$ ) and a per-user hit rate of 20.5% ( $\sigma=16.4$ ).

### 4.3 Temporal Patterns of Genre Selection

Data from 456 users, including more than 5100 songs and 117 genres, were used for the validation of the genre-related patterns. In order to consider a "hit" in the prediction of listening time and day for a given genre, we set the same thresholds than for evaluating the artist prediction. For the time of the day an overall 22.6% (and per-user 23.2%) of accurate predictions was found. It is interesting to note that relaxing the required accuracy of the prediction to plus/minus one hour error we reached 39.9% of

average hits and per-user average 41% ( $\sigma=28.4$ ). For the day of the week, the overall hit percent was 40.9%, while the per-genre average and the per-user average were, respectively, 40.7% ( $\sigma=24.1$ ) and 41.7% ( $\sigma=26.3$ ). It is interesting to note that among the best predictable genres we find many of infrequent ones but also many of the most frequent ones.

## 6. CONCLUSIONS

The present study is, as far as we know, the first one inquiring the possibility that our music listening behavior may follow some detectable circadian and ultradian patterns, at least under certain circumstances. We have discovered that a non-negligible amount of listeners tend to prefer to listen to certain artists and genres at specific moments of the day and/or at certain days of the week. We have also observed that, respectively for artists and for genres, 20% and 40% time-contextualized music recommendations can be successful. In our future work agenda, more sophisticated prediction models will be tested, and also ways to implement them into existing music recommenders.

## 7. ACKNOWLEDGMENTS

Our thanks to Óscar Celma who kindly shared the Last.fm data file, accessible from this URL:

<http://www.dtic.upf.edu/~ocelma/MusicRecommendationDataset/lastfm-1K.html>

## 8. REFERENCES

- [1] Anderson, M., Ball, M., Boley, H., Greene, S., Howse, N., Lemire, D. and McGrath, S. 2003. Racofi: A rule-applying collaborative filtering system. In Proc. of COLA'03.
- [2] Baltrunas, L. and Amatriain, X. 2009. Towards Time-Dependant recommendation based on implicit feedback. RecSys09 Workshop on Context-aware Recommender Systems (CARS-2009).
- [3] Balzer, H.U. 2009. Chronobiology as a foundation for and an approach to a new understanding of the influence of music. In R. Haas and V. Brandes (Eds.), *Music that Works*. Wien/New York: Springer Verlag.
- [4] Barabasi, A.L. 2010. *Bursts: The Hidden Pattern Behind Everything We Do*. New York: Dutton Books.
- [5] Beran, J. 2004. *Statistics in Musicology*, Boca Raton: CRC.
- [6] Berens P., 2009, CircStat, a Matlab Toolbox for Circular Statistics, *Journal of Statistical Software*, 31, 10.
- [7] Boström, F. 2008. AndroMedia - Towards a Context-aware Mobile Music Recommender. Master's thesis, University of Helsinki, Faculty of Science, Department of Computer Science. <https://oa.doria.fi/handle/10024/39142>.
- [8] Coppola, P., Della Mea, V., Di Gaspero, L., Menegon, D., Mischis, D., Mizzaro, S., Scagnetto, I. and Vassena, L. 2009. The context-aware browser. *IEEE Intelligent Systems*, 25,1, 38-47.
- [9] Dressler, K. and Streich, S. 2007. Tuning Frequency Estimation Using Circular Statistics. 8<sup>th</sup> Int. Conf. on Music Information Retrieval (ISMIR-2007), 357-360.
- [10] Eagle, N. and Pentland, A.S. 2009. Eigenbehaviors: Identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63, 7, 1057-1066.
- [11] Fisher N.I., 1993, *Statistical Analysis of circular data*, Cambridge: Cambridge University Press.
- [12] Foster, R.G., and Kreitzman, L. 2005. *Rhythms of Life: The Biological Clocks that Control the Daily Lives of Every Living Thing*. Yale: Yale University Press.
- [13] Hargreaves, D. J. and North, A. C. 1999. *The functions of music in everyday life: Redefining the social in music psychology*. *Psychology of Music* 27, 71-83.
- [14] Koren, Y. 2009. Collaborative filtering with temporal dynamics, New York, NY, USA, 447-456.
- [15] Kubiak, T. and Jonas, C. 2007. Applying circular statistics to the analysis of monitoring data: Patterns of social interactions and mood. *European Journal of Personality Assessment*, 23, 227-237.
- [16] Larsen, R.J., Augustine, A.A., and Prizmic, Z. 2009. A process approach to emotion and personality: Using time as a facet of data. *Cognition and Emotion*, 23, 7, 1407-1426.
- [17] Lee, J.S. and Lee, J.C. 2008. Context awareness by case-based reasoning in a music recommendation system. 4<sup>th</sup> Int. Conf. on Ubiquitous Computing Systems, 45-58.
- [18] Lloyd, D., and Rossi, E. 2008. *Ultradian Rhythms from Molecules to Mind: a new vision of life*. New York: Springer.
- [19] Lombardi, S., Anand, S. and Gorgoglione, M. 2009. Context and Customer Behavior in Recommendation. RecSys09 Workshop on Context-aware Recommender Systems.
- [20] Neuhaus, F., 2010. *Cycles in Urban Environments: Investigating Temporal Rhythms*. Saarbrücken: LAP.
- [21] Radocy, R.E. and Boyle, J.D. 1988. *Psychological Foundations of Musical Behavior* (2nd ed.) Springfield, IL: Charles C. Thomas.
- [22] Rentfrow, P.J. and Gosling, S.D. 2003. The do re mi's of everyday life: The structure and personality correlates of music preferences. *Journal of Personality and Social Psychology*, 84, 6, 1236-1256.
- [23] Reynolds, G., Barry, D., Burke, T., and Coyle, E. 2008. Interacting with large music collections: towards the use of environmental metadata. *IEEE International Conference on Multimedia and Expo*, 989-992.
- [24] Saarikallio, S., and Erkkilä, J. 2007. The role of music in adolescents' mood regulation. *Psych. of Music*, 35, 1, 88-109.
- [25] Su, J.H., Yeh, H.H., Yu, P.S., Tseng, V. 2010. Music recommendation using content and context information mining. *IEEE Intelligent Systems*, 25, 16-26.
- [26] Valcheva, M. 2009. Playlistism: a means of identity expression and self-representation. Technical Report, Intermedia, University of Oslo. [http://www.intermedia.uio.no/download/attachments/43516460/vit-ass-mariya\\_valcheva.pdf?version=1](http://www.intermedia.uio.no/download/attachments/43516460/vit-ass-mariya_valcheva.pdf?version=1)
- [27] Zar J.H. 1999, *Biostatistical Analysis* (4<sup>th</sup> edition), Upper Saddle River, NJ: Prentice Hall.



## 5. THE AGE OF CREATIVE SYSTEMS?

*In the future, you won't buy artists' works;  
you'll buy software that makes original pieces of "their" works,  
or that recreates their way of looking at things.*

Brian Eno, Wired 3.05, May 1995, p. 150

### 5.1. Introduction

In the preceding decades, we have witnessed how the interest on and the needs of analysing, describing, and experiencing music contents, motivated research efforts along three sequentially opened but intercommunicated and coexistent paths that we have titled “ages”. Currently available music representations are rich (even “deep”) though they are still incomplete, not only because of technical limitations in the available algorithms and data, but also because one essential component was frequently ignored. Music is, above all, social action: performance happens before listening, before analysing, individually or collectively. Performance predates millenia any known form of symbolic record related to music. And music performance, additionally, is originally social. It tends to be created between people that, sharing some framework, interact under accepted codes to create music. It is then natural to wonder if the performative and interactive aspects of music can be benefited from state-of-the-art techniques and technologies developed in twenty years of MIR. Moreover, a digital native generation, used to constant interaction with devices as proxies of people, places, emotions or actions, would acclaim any development where information extracted from music items could be put into value by giving them the agency to move it around, change it, alter other processes, etc.

When we use the expression “creative systems” there are two possible senses for that concept: either systems that show creativity by themselves (hence, artificial systems with a certain autonomy in the decisions they make), or systems that enhance creativity of humans (hence, tools for creators who are -or should be- the ultimate decision makers). It cannot be denied that humans show a bizarre fascination for any kind of automatism, and automatic music creation systems have been around us since the invention of the clockwork (or probably before, at least as “algorithms” -what, if not an algorithm, is implicit in a treatise on harmonization or on counterpoint? -). In these and many other algorithm composition systems (Hiller & Isaacson, 1959; Cope, 1991) some arbitrary rules have been implemented to generate music with certain particularities (i.e., those defining fifth species counterpoint, for example). Currently, though, we do not need to be restricted to deal with such low-interaction systems and we can get more insights and satisfactions (as researchers and as users too) with those of the other type, those that allow

creation to develop as a human-machine dialog, turn-taking, negotiation... processes that humans use in a natural way<sup>35</sup>.

It is significant that, in ISMIR 2013, a late-breaking demo session was devoted to “Creative MIR” (Humphrey, Turnbull and Collins, 2013). Research on that field had already been presented in earlier conference editions or elsewhere (Wiggins, 2006; Griffin et al., 2010; Fiebrink, 2011; Pardo et al., 2012) but this could be taken as the implicit acknowledgment of a new “age” starting. In addition, I was recently surprised to discover in the news section of the Computer Music Journal (Unknown reporter, 2011) a very short summary of my keynote to the Third International Workshop on Advances in Music Information Research (Ad-MIRe 2011) putting in my mouth the words “time has arrived for a paradigm shift towards doing use-inspired basic research where the focus on ‘information’ shifts towards ‘interaction’”<sup>36</sup>. So, then we were all tuned-in, it seems. Now features, semantic content and context can be flowing to assist music creation, and some of the to-be included elements are gestures, haptics and a closed loop between the user and the system. In this context was conceived and developed the GiantSteps project<sup>37</sup>, which started in 2013, and provided the context where the paper included in this chapter originated.

The MIREs<sup>38</sup> roadmap (Serra et al., 2013) was already remarking the growing importance of MIR for music creation, with the goal of researching on “intuitive tools, controlled through parameters relevant from the viewpoint of human cognition of music and sound, and also to enhance the quality of existing processes by selecting appropriate processes and parameter sets according to the nature of the extracted elements”. Some areas that are considered ripe for that are:

- Content-based sound processing: commercial examples of using audio analysis outcomes to guide sound edition and mixing can be experienced with GarageBand, Melodyne, Revoice Pro or Ableton Live, which use polyphonic transcription, drum detection, or tempo and pitch adaptation to input or guide materials. Semantic-based (or automatic) mixing is another sparkling and promising area (Man & Reiss, 2013; Man et al., 2017). Lastly, let’s mention the innovative possibility of object and content-based processing (Hattwick et al., 2013) whereby objects capable of sensing and analysing sound and gestures foster original ways to musically interact with the environment (see also Mogeess<sup>39</sup>).
- Computer-aided composition. In contrast to more traditional and general-purpose systems, the possibility of modelling styles and genres is one of the most thrilling avenues for research because it calls for the analysis and representation of specific corpora, sometimes in an interactive fashion. Systems that, with a minimal input from the user, can generate reasonable style-constrained variations (Henry et al., 2018) or complement such input (i.e., by harmonizing with it, by adding a bassline, generating solos...), are demanded by potential users (Andersen and Knees, 2016), have been early research topics (Arcos et al, 1998 Grachten et al.,

---

<sup>35</sup> See Herrerman et al. (2017) for an excellent review and systematization of music generation systems.

<sup>36</sup> The keynote title was “From MIR to MIR Through Three Ages and a Paradigm Shift”, and there I presented some of the ideas that have been framed this thesis

<sup>37</sup> <http://www.giantsteps-project.eu/#/>

<sup>38</sup> <http://www.mires.cc/>, but see also <http://mtg.upf.edu/MIREs>.

<sup>39</sup> <https://www.facebook.com/mogeess/>

2006), and they still pose challenging research questions (McVicar et al., 2014; Van Balen, 2016; Faraldo et al., 2017; Gómez-Marín et al, 2017). The project Flow Machines has fuelled the field with surprising and provocative outcomes<sup>40</sup> but also with relevant research contributions, many of them related to constrained Markov models (Pachet et al., 2013; Roy et al, 2016; Sakellariou et al., 2017).

- Databases for music and sound production, where the prototypical applications are content-driven concatenative synthesis (Zils & Pachet, 2001; Collins, 2002; Schwartz et al., 2006) mashup generation (Davies et al., 2014), or soundtrack generation (Müller & Driedger, 2012). Several systems have been developed more recently such as LoopMashVST<sup>41</sup>, FreeSound Radio (Roma et al., 2009), Floop (Roma & Serra, 2015), Phonos Explorer (Bandiera, 2015), FreeSound Explorer (Font et al., 2017), and APICultor (Ordiales and Bruno, 2017). In 2018 Native Instruments has opened a web-based sample-search service powered with MIR algorithms<sup>42</sup> (it is worth to note that 15 years have passed since IRCAM's Studio Online (Ballet & Borghessi, 1999) or our AUDIOCLAS demonstrators for the Tape Gallery (see chapter 3), which essentially were offering a similar service, even though the technologies were more rudimentary and error-prone).
- Live performance spans through different and usually separated areas. In the context of DJing, for example, beat synchronisation and harmonic mixing (Ishizaki, 2015), are typical problems. Here we have witnessed the development of virtual DJs (Cliff, 2000; Hirai, 2015), thanks to the work done on automatic playlist generation and continuation. The problem even motivated a Turing test competition<sup>43</sup> (in which one of the winners (Parera, 2016) was developed as a MSc thesis project supervised by Sergi Jordà and the author of this thesis<sup>44</sup>), but it seems much more interesting and full-fledged the AI DJ project<sup>45</sup>, even though it has not been academically presented. With quite a different framework and aesthetics, live coding is another approach that keeps growing, with proposals that, for example, combine live coding and the analysis of personal or public collections of sounds (Xambó et al., 2018). The work developed in the IRCAM around OMax and based on audio-based oracle factors (Assayag et al., 2010; Dubnov & Assayag, 2012), makes possible the establishment of improvisatory dialogues between humans and music-listening (music recycling) systems or, additionally, interactive arrangement and voicing based on pre-set corpuses. Also based on oracle factors is Mimi (Schankler, 2014), a system that allows the visualization of the musical past and the (potential) futures of a live improvisation. Another popular performance technique, looping, has also been addressed in the context of Flow Machines (Pachet et al., 2013).

---

<sup>40</sup> <http://www.flow-machines.com/overview/>,  
[https://www.youtube.com/channel/UCB00CfzP5YHpgJbMB\\_exm9A](https://www.youtube.com/channel/UCB00CfzP5YHpgJbMB_exm9A)

<sup>41</sup> <https://www.soundonsound.com/techniques/great-loops-loopmash>,  
[https://www.steinberg.net/en/products/mobile\\_apps/loopmash.html](https://www.steinberg.net/en/products/mobile_apps/loopmash.html)

<sup>42</sup> <https://sounds.com/>

<sup>43</sup> <http://bregman.dartmouth.edu/turingtests/music2018>. This competition has changed the goal in subsequent years, dealing now with style imitation and improvisation with a human performer.

<sup>44</sup> <http://bregman.dartmouth.edu/turingtests/node/57>

<sup>45</sup> <https://medium.com/@naotokui/ai-dj-project-a-dialog-between-human-and-ai-through-music-abca9fd4a45d>

The age of creative music systems takes advantage of the hybridization of several research communities, specially those of MIR, AI and HCI. Concepts such as “The Machine Learning Algorithm as Creative Musical Tool” (Fiebrink & Caramiaux, 2018) are keeping our expectations higher than ever.

## 5.2. Papers included in this chapter

Nuanáin, C. Ó., **Herrera P.**, & Jordà S. (2017). Rhythmic Concatenative Synthesis for Electronic Music: Techniques, Implementation, and Evaluation. *Computer Music Journal*. 41(2), 21-37 (Journal h-index: 35; Journal IF 2016: 0.405; Q1 in Music Journals; 0 citations; a shorter version was selected best paper in NIME 2016; CMJ used a screenshot for the cover of the issue where it was published).

## 5.3. Contributions in the selected papers

In addition to our contribution of providing a quite complete state of the art on concatenative synthesis and its application and relevance in the composition and production of styles of electronic dance music, we describe RhythmCAT, a user-friendly system for generating rhythmic loops that model the timbre and rhythm of an initial target loop. Although there are a few commercial applications that encapsulate techniques of concatenative synthesis for the user, most systems are custom-built for the designer or are prototypical in nature. Additionally, we detected a marked lack of evaluation strategies or reports of user experiences in the accompanying literature and, based on these investigations, we set out to design a system that applied and extended many of the pervasive techniques in concatenative synthesis with a clear idea of its application and its target user. We built then an instrument that was easily integrated with modern digital audio workstations and presented an interface that intended to be attractive and easy to familiarize oneself with. The system operates as an Ableton Live instrument that takes an audio excerpt or a loop as input and “reconstructs” it using a provided collection of sound samples. The system makes possible to use a personal similarity function, assigning different weights to the involved features. One of the most appealing features is a 2D interactive timbre space where users can modulate, in real-time, the concatenation sequence.

A three-tiered, qualitative and quantitative, evaluation of the system, not only in terms of its objective performance but also in the subjective aural and experiential implications for our users, was our final substantial contribution to this area. The results of our evaluations showed that our system is an efficient, effective and user-friendly tool (though users were not favourable to personalising the similarity function, for example) that integrates well in the typical workflow of electronic music creators.

Nuanáin, C. Ó., **Herrera P.**, & Jordà S. (2017). Rhythmic Concatenative Synthesis for Electronic Music: Techniques, Implementation, and Evaluation. *Computer Music Journal*. 41(2), 21-37.

[DOI; https://doi.org/10.1162/COMJ\\_a\\_00412](https://doi.org/10.1162/COMJ_a_00412)

ISSN: 0148-9267

Online ISSN: 1531-5169



---

**Cárthach Ó Nuanáin, Perfecto  
Herrera, and Sergi Jordà**

Music Technology Group  
Communications Campus–Poblenou  
Universitat Pompeu Fabra  
Carrer Roc Boronat, 138, 08018  
Barcelona, Spain  
{carthach.onuanain, perfecto.herrera,  
sergi.jorda}@upf.edu

## Rhythmic Concatenative Synthesis for Electronic Music: Techniques, Implementation, and Evaluation

**Abstract:** In this article, we summarize recent research examining concatenative synthesis and its application and relevance in the composition and production of styles of electronic dance music. We introduce the conceptual underpinnings of concatenative synthesis and describe key works and systematic approaches in the literature. Our system, RhythmCAT, is proposed as a user-friendly system for generating rhythmic loops that model the timbre and rhythm of an initial target loop. The architecture of the system is explained, and an extensive evaluation of the system's performance and user response is discussed based on our results.

Historically, reusing existing material for the purposes of creating new works has been a widely practiced technique in all branches of creative arts. The manifestations of these expressions can be wholly original and compelling, or they may be derivative, uninspiring, and potentially infringe on copyright (depending on myriad factors including the domain of the work, the scale of the reuse, and cultural context).

In the visual arts, reusing or adapting existing material is most immediately understood in the use of collage, where existing works or parts thereof are assembled to create new artworks. Cubist artists such as Georges Braque and Pablo Picasso extensively referenced, appropriated, and reinterpreted their own works and the works of others, as well as common found objects from their surroundings (Greenberg 1971). Collage would later serve as direct inspiration for bricolage, reflecting wider postmodernist trends towards deconstructionism, self-referentiality, and revisionism that include the practice of parody and pastiche (Lochhead and Auner 2002).

In music and the sonic arts, the natural corollary of collage came in the form of *musique concrète* (Holmes 2008), a movement of composition stemming from the experiments of Pierre Schaeffer and, later, Pierre Henry at the studios of Radiodiffusion-Télévision Française in Paris during the 1940s and 1950s (Battier 2007). In contrast to the artificially

and electronically generated *elektronische Musik* spearheaded by Karlheinz Stockhausen at the West German Radio studios in Cologne, the French composers sought to conceive their works from existing recorded sound, including environmental sources like trains and speech. Seemingly unrelated and nonmusical sounds are organized in such a way that the listener discovered the latent musical qualities and structure they inherently carry.

It is important to note that in music composition general appropriation of work predates these electronic advancements of technology. In Western art music, for example, composers like Béla Bartók—himself a musicologist—have often turned to folk music for its melodies and dance music styles (Bartók 1993), and others (e.g., Claude Debussy, cf. Tamagawa 1988) became enchanted by music from other cultures, such as Javanese gamelan, studying its form and incorporating the ideas into new pieces. Quotations, or direct lifting of melodies from other composers' works, are commonplace rudiments in jazz music. Charlie Parker, for example, was known to pepper his solos with reference to Stravinsky's *Rite of Spring* (Mangani, Baldizzone, and Nobile 2006). David Metzger has compiled a good reference on appropriation and quotation music (Metzger 2003).

The modern notion of sampling stems from the advent of the digital sampler and its eventual explosion of adaptation in hip-hop and electronic music. Artists such as Public Enemy and the Beastie Boys painstakingly assembled bewildering permutations of musical samples, sound bites, and other miscellaneous recorded materials that sought to supplant

---

the many cultural references that permeated their lyrics (Sewell 2014). Later, the influence of hip-hop production would inform the sample-heavy arrangements of jungle and drum and bass, in particular with its exhaustive rerendering of the infamous “Amen Break.” John Oswald, an artist who directly challenged copyright for artistic gain, dubbed his approach “plunderphonics” and set out his intentions in a suitably subtitled essay “Plunderphonics, or Audio Piracy as a Compositional Prerogative” (Oswald 1985). Using tape-splicing techniques, he created deliberately recognizable montages of pop music, such as that by Michael Jackson, in a style that became later known as “mashups.” Nowadays, artists such as Girtalk create extremely complex and multireferential mashups of popular music, harnessing the powerful beat-matching and synchronization capabilities of the modern digital audio workstation (Humphrey, Turnbull, and Collins 2013).

Although the question of originality and authorship is not in the realm of this discussion, this interesting and pertinent topic is under the scrutiny of researchers in musicology and critical studies. We encourage the reader to consult work by Tara Rodgers (2003), Paul Miller (2008), and Kembrew McLeod (2009) for a more focused discourse.

Associated research efforts in computer music, signal processing, and music information retrieval (MIR) afford us the opportunity to develop automated and intelligent systems that apply the aesthetic of sampling and artistic reuse. The term *concatenative synthesis* has been extensively used to describe musical systems that create new sound by automatically recycling existing sounds according to some well-defined set of criteria and algorithmic procedures. Concatenative synthesis can be considered the natural heir of granular synthesis (Roads 2004), a widely examined approach to sound synthesis using tiny snippets (“grains”) of around 20–200 msec of sound, which traces its history back to Iannis Xenakis’s theories in *Formalized Music* (Xenakis 1971). With concatenative synthesis, the grains become “units” and are more related to musical scales of length, such as notes and phrases. Most importantly, information is attached to these units of sound: crucial descriptors that allow spectral and temporal characteristics

of the sound to determine the sequencing of final output.

In the following sections, we will present a thorough, critical overview of many of the key works in the area of concatenative synthesis, based on our observation that there has not been such a broad survey of the state of the art in other publications in recent years. We will compare and contrast characteristics, techniques, and the challenges of algorithmic design that repeatedly arise. For the past three years, we have been working on the European-led initiative GiantSteps (Knees et al. 2016). The broad goal of the project is the research and development of expert agents for supporting and assisting music makers, with a particular focus on producers of electronic dance music (EDM). Consequently, one of the focuses of the project has been on user analysis: thinking about their needs, desires, and skills; investigating their processes and mental representations of tasks and tools; and evaluating their responses to prototypes.

Modern EDM production is characterized by densely layered and complex arrangements of tracks making liberal use of synthesis and sampling, exploiting potentially unlimited capacity and processing in modern computer audio systems. One of our main lines of research in this context has been the investigation of concatenative synthesis for the purposes of assisting music producers to generate rhythmic patterns by means of automatic and intelligent sampling.

In this article, we present the RhythmCAT system, a digital instrument that creates new loops emulating the rhythmic pattern and timbral qualities of a target loop using a separate corpus of sound material. We first proposed the architecture of the system in a paper for the conference on New Interfaces for Musical Expression (Ó Nuanáin, Jordà, and Herrera 2016a), followed by papers evaluating it in terms of its algorithmic performance (Ó Nuanáin, Herrera, and Jordà 2016) and a thematic analysis of users’ experience (Ó Nuanáin, Jordà, and Herrera 2016b). This article thus represents an expanded synthesis of the existing literature, our developments motivated by some detected shortcomings, and the illustration of an evaluation strategy.



---

## State of the Art in Concatenative Synthesis

Other authors have previously provided insightful summaries of research trends in concatenative synthesis (e.g., Schwarz 2005; Sturm 2006). These surveys are over ten years old, however (but see Schwarz 2017 for a continuously updated online survey), so we offer here a more recent compendium of state-of-the-art systems as we see them, based on our investigations of previous publications up until now.

Before music, concatenative synthesis enjoyed successful application in the area of speech synthesis; Hunt and Black (1996) first reported a unit selection scheme using hidden Markov models (HMMs) to automatically select speech phonemes from a corpus and combine them into meaningful and realistic sounding sentences. Hidden Markov models extend Markov chains by assuming that “hidden” states output visible symbols, and the Viterbi algorithm (Rabiner 1989) can return the most probable sequence of states given a particular sequence of symbols. In concatenative synthesis, the maximum probabilistic model is inverted to facilitate minimal cost computations.

The *target cost* of finding the closest unit in the corpus to the current target unit becomes the emission probability, with the *concatenation cost* representing the transition probability between states. The Viterbi algorithm thus outputs indices of database units corresponding to the optimal state sequence for the target, based on a linear combination of the aforementioned costs. Diemo Schwarz (2003) directly applied this approach for musical purposes in his Caterpillar system.

Schwarz notes, however, that the HMM approach can be quite rigid for musical purposes because it produces one single optimized sequence without the ability to manipulate the individual units. To address these limitations, he reformulates the task into a constraint-satisfaction problem, which offers more flexibility for interaction. A constraint-satisfaction problem models a problem as a set of variables, values, and a set of constraints that allows us to identify which combinations of variables and values are violations of those constraints, thus allowing us to quickly reduce large

portions of the search space (Russell and Norvig 2009).

Zils and Pachet (2001) first introduced constraint satisfaction for concatenative synthesis in what they describe as musical mosaicking—or, to use their portmanteau, *musicing*. They define two categories of constraints: *segment* and *sequence* constraints. Segment constraints control aspects of individual units (much like the target cost in an HMM-like system) based on their descriptor values. Sequence constraints apply globally and affect aspects of time, continuity, and overall distributions of units. The constraints can be applied manually by the user or learned by modeling a target. The musically tailored “adaptive search” algorithm performs a heuristic search to minimize the total global cost generated by the constraint problem. One immediate advantage of this approach over the HMM is the ability to run the algorithm several times to generate alternative sequences, whereas the Viterbi process always outputs the most optimal solution.

A simpler approach is presented in MatConcat (Sturm 2004), using feature vectors comprising six descriptors and computing similarity metrics between target units and corpus units. Built for the MATLAB environment for scientific computing, the interface is quite involved, and the user has control over minute features such as descriptor tolerance ranges, relative descriptor weightings, as well as window types and hop sizes of output transformations. On Sturm’s Web site are short compositions generated by the author using excerpts from a Mahler symphony as a target, and resynthesized using various unrelated sound sets, for instance, pop vocals, found sounds, and solo instrumental recordings from saxophone and trumpet ([www.mat.ucsb.edu/~b.sturm/music/CVM.htm](http://www.mat.ucsb.edu/~b.sturm/music/CVM.htm)).

As concatenative synthesis methods matured, user modalities of interaction and control became more elaborate and real-time operations were introduced. One of the most compelling features of many concatenative systems is the concept of the *interactive timbre space*. With the release of CataRT (Schwarz et al. 2006), these authors provided an interface that arranges the units in an interactive two-dimensional timbre space. The arrangement

---

of these units is according to a user-selectable descriptor on each axis. Instead of using a target sound file to inform the concatenation procedure, the user's mouse cursor becomes the target. Sounds that are within a certain range of the mouse cursor are sequenced according to some triggering options (one-shot, loop, and—most crucially—with real-time output).

Bernardes takes inspiration from CataRT and from Tristan Jehan's Skeleton (Jehan 2005) to build his EarGram system for the Pure Data (Pd) environment (Bernardes, Guedes, and Pennycook 2013). Built on top of William Brent's excellent feature-extraction library timbreID (Brent 2010), it adds a host of interesting features for visualization and classification. For instance, as well as the familiar waveform representation and previously described 2-D timbre representation (with various clustering modes and dimensionality-reduction implementations), there are similarity matrices that show the temporal relations in the corpus over time. Some unique playback and sequencing modes also exist, such as the *infiniteMode*, which generates endless playback of sequences, and the *soundscapeMap*, which features an additional 2-D control of parameters pertaining to sound scene design. Another system that adapts a 2-D timbre space is AudioGarden by Frisson, Picard, and Tardieu (2010), which offers two unique mapping procedures. The first of these, "disc" mode, places units by assigning the length of the audio file to the radius of the unit from the center, with the angle of rotation corresponding to a principal component of timbre, mel-frequency cepstrum coefficients (MFCCs). In the other mode, called "flower" mode, a point of the sound is positioned in the space according to the average MFCCs of the entire sound file. Segments of the particular sound are arranged in chronological fashion around this center point.

There have been some concatenative systems tailored specifically with rhythmic purposes in mind. Pei Xiang proposed Granuloop for automatically re-arranging segments of four different drum loops into a 32-step sequence (Xiang 2002). Segmentation is done manually, without the aid of an onset detector, using the Recycle sample editor from Propellerhead Software. Segmented sounds are compared using the

inner product of the normalized frequency spectrum, supplemented with the weighted energy. These values become weights for a Markov-style probability transition matrix. Implemented in Pd, the user interacts by moving a joystick in a 2-D space, which affects the overall probability weightings determining which loop segments are chosen. The system presents an interesting approach but is let down by its lack of online analysis. Ringomatic (Aucouturier and Pachet 2005) is a real-time agent specifically tailored for combining drum tracks, expanding on many of the constraint-based ideas from their prior musaicing experiments. They applied the system to real-time performance following symbolic feature data extracted from a human MIDI keyboard player. They cite, as an example, that a predominance of lower-register notes in the keyboard performance applies an inverse constraint that creates complementary contrast by specifying that high-frequency heavy cymbal sounds should be concatenated.

As demonstrated in EarGram, concatenative synthesis has been considered useful in sound design tasks, allowing the sound designer to build rich and complex textures and environments that can be transformed in many ways, both temporally and timbrally. Cardle, Brooks, and Robinson (2003) describe their Directed Sound Synthesis software as a means of providing sound designers and multimedia producers a method of automatically reusing and synthesizing sound scenes in video. Users select one or more regions of an existing audio track and can draw probability curves on the timeline to influence resynthesis of these regions elsewhere (one curve per region). Hoskinson and Pai (2001), in a nod to granular synthesis, refer to the segments used in their Soundscapes software as "natural grains," and they seek to synthesize endless streams of soundscapes. The selection scheme by which segments are chosen is based on a representation of each segment as a transition state in a Markov chain. Its interface features knobs and sliders for interactively controlling gain and parameters of multiple samples. To evaluate the platform they conducted an additional study (Hoskinson and Pai 2007) to reveal whether listening subjects found the concatenated sequences convincing compared with genuinely recorded soundscapes.

Figure 1. Block diagram of functionality in the RhythmCAT system.

More-specific and applied-use cases of concatenative synthesis include work by Ben Hackbarth, who explores the possibilities of concatenative synthesis in large-scale music composition (Hackbarth, Schnell, and Schwarz 2011). Hackbarth has worked with Schwarz to provide an alternative interface for exploring variations based on a force-directed graph. John O’Connell describes a graphical system for Pd that demonstrates the use of higher-level perceptual concepts like mood (happy versus sad) for informing selection in audio mosaics (O’Connell 2011).

Commercial implementations also exist for concatenative synthesis. Of particular note is Steinberg’s Loopmash, a software plug-in and mobile application for automatically creating mashups from existing looped content ([www.steinberg.net/loopmash](http://www.steinberg.net/loopmash)). The interface consists of a number of tracks in a timeline arrangement. One track is set as a master, and slices in the master are replaced with matching slices from the other slave tracks. Users interact by manipulating “similarity gain” sliders that control the influence of each track in the slice selection algorithm. Other applications exist more as MIDI sampler systems attempting to model the performance qualities of natural sources such as orchestral ensembles (e.g., SynfulOrchestra, [www.synful.com](http://www.synful.com)) or the human voice (e.g., Vocaloid, [www.vocaloid.com](http://www.vocaloid.com)).

There are many other concatenative systems that are too numerous to discuss in detail here. We have, however, compiled a table in a previous publication summarizing all the systems we have come across in our research, with remarks on interaction and visualization features, support for rhythm, and whether any user evaluation was carried out (Ó Nuanáin, Jordà, and Herrera 2016b).

## Design and Implementation

In this section, we will describe our implementation of the RhythmCAT system, beginning with an explanation of the musical analysis stages of onset detection, segmentation, and feature extraction. This is followed by an examination of the interactive user interface and the pattern-generation process.

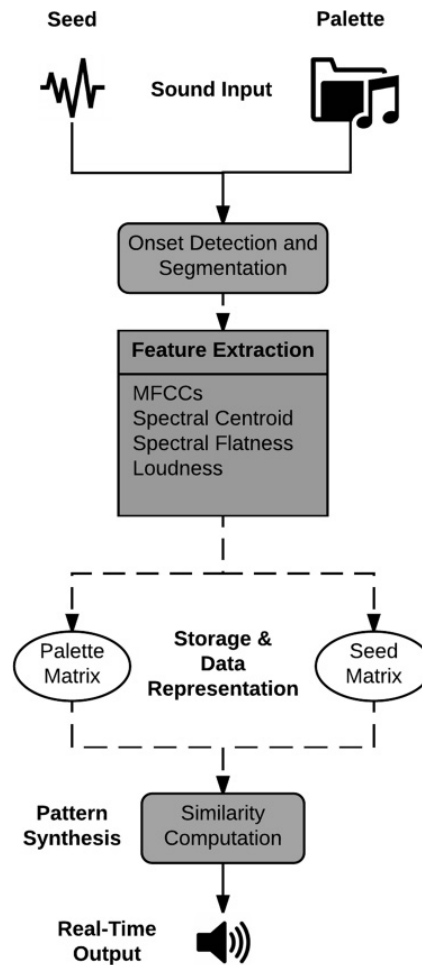


Figure 1 gives a diagrammatic overview of these important stages, which can be briefly summarized as:

1. Sound Input
2. Onset Detection and Segmentation
3. Audio Feature Extraction
4. Storage and Data Representation
5. Pattern Synthesis
6. Real-Time Audio Output

---

The system is developed in C++ using the JUCE framework ([www.juce.com](http://www.juce.com)), the Essentia musical analysis library (Bogdanov et al. 2013), and the OpenCV computer vision library for matrix operations (Bradski 2000).

### Sound Input

The first stage in building a concatenative music system generally involves gathering a database of sounds from which selections can be made during the synthesis procedure. This database can be manually assembled, but in many musical cases the starting point is some user-provided audio that may range in length from individual notes to phrases to complete audio tracks.

The two inputs to the system are the *sound palette* and the *seed sound*. The sound palette refers to the pool of sound files we want to use as the sample library for generating our new sounds. The seed sound refers to the short loop that we wish to use as the similarity target for generating those sounds. The final output sound is a short (one to two bars) loop of concatenated audio that is rendered in real time to the audio host.

### Onset Detection and Segmentation

In cases where the sounds destined for the sound palette exceed note or unit length, the audio needs to be split into its constituent units using onset detection and segmentation.

Onset detection is a large topic of continuous study, and we would encourage the reader to examine the excellent review of methods summarized by Simon Dixon (2006). Currently, with some tuning of the parameters, Sebastien Bock's Superflux algorithm represents one of the best-performing state-of-the-art detection methods (Böck and Widmer 2013). For our purposes, we have experienced good results with the standard onset detector available in Essentia, which uses two methods based on analyzing signal spectra from frame to frame (at a rate of around 11 msec). The first method involves estimating the high-frequency content in each frame

(Masri and Bateman 1996) and the second method involves estimating the differences of phase and magnitude between each frame (Bello and Daudet 2005).

The onset detection process produces a list of onset times for each audio file, which we use to segment into new audio files corresponding to unit sounds for our concatenative database.

### Audio Feature Extraction

In MIR systems, the task of deciding which features are used to represent musical and acoustic properties is a crucial one. It is a trade-off between choosing the richest set of features capable of succinctly describing the signal, on the one hand, and the expense of storage and computational complexity, on the other. When dealing specifically with musical signals, there are a number of standard features corresponding roughly to certain perceptual sensations. We briefly describe the features we chose here (for a more thorough treatment of feature selection with relation to percussion, see Herrera, Dehamel, and Gouyon 2003; Tindale et al. 2004; and Roy, Pachet, and Krakowski 2007).

Our first feature is the loudness of the signal, which is implemented in Essentia according to Steven's Power Law, namely, the energy of the signal raised to the power of 0.67 (Bogdanov et al. 2013). This is purported to be a more perceptually effective measure for human ears. Next, we extract the spectral centroid, which is defined as the weighted mean of the spectral bins extracted using the Fourier transform. Each bin is then weighted by its magnitude.

Perceptually speaking, the spectral centroid relates mostly to the impression of the brightness of a signal. In terms of percussive sounds, one would expect the energy of a kick drum to be more concentrated in the lower end of the spectrum and hence have a lower centroid than that from a snare or crash cymbal.

Another useful single-valued spectral feature is the spectral flatness. It is defined as the geometric mean of the spectrum divided by the arithmetic mean of the spectrum. A spectral flatness value of 1.0

---

means the energy spectrum is flat, whereas a value of 0.0 would suggest spikes in the spectrum indicating harmonic tones (with a specific frequency). The value intuitively implies a discrimination between noisy or inharmonic signals and signals that are harmonic or more tonal. Kick-drum sounds (especially those generated electronically) often comprise quite a discernible center frequency, whereas snares and cymbals are increasingly broadband in spectral energy.

Our final feature is MFCCs. These can be considered as a compact approximation of the spectral envelope and is a useful aid in computationally describing and classifying the timbre of a signal. It has been applied extensively in speech processing, genre detection (Tzanetakis, Essl, and Cook 2001), and instrument identification (Loughran et al. 2004). The computation of MFCCs, as outlined by Beth Logan (2000), is basically achieved by computing the spectrum, mapping the result into the more perceptually relevant mel scale, taking the log, and then applying the discrete cosine transform.

It is difficult to interpret exactly what each of the MFCC components mean, but the first component is generally regarded as encapsulating the energy. Because we are already extracting the loudness using another measure, we have discarded this component in our system. For detailed explanations and formulae pertaining to the features introduced here, as well as others, we direct the reader to Geoffroy Peeters's compendium (Peeters 2004).

### Storage and Data Representation

Further on in this article we will describe in greater detail how the seed or target audio signal is actually received from the Virtual Studio Technology host, but in terms of analysis on that seed signal, the process is the same as before: onset detection and segmentation followed by feature extraction.

The resulting feature vectors are stored in two matrices: the palette matrix and the target matrix. The palette matrix stores the feature vectors of each unit of sound extracted from the sound palette, and the target matrix similarly stores feature vectors of units of sound extracted from the seed loop.

### Pattern Synthesis and Real-Time Audio Output

This section details the visible, aural, and interactive elements of the system as they pertain to the user. Figure 2 provides a glimpse of the user interface in a typical pattern generation scenario.

#### Workflow

The layout of the interface was the result of a number of iterations of testing with users who, while praising the novelty and sonic value of the instrument, sometimes expressed difficulty understanding the operation of the system. One of the main challenges faced was how best to present the general workflow to the user in a simple and concise manner. We decided to represent the flow of the various operations of the software emphatically by using a simple set of icons and arrows, as seen in Figure 2a.

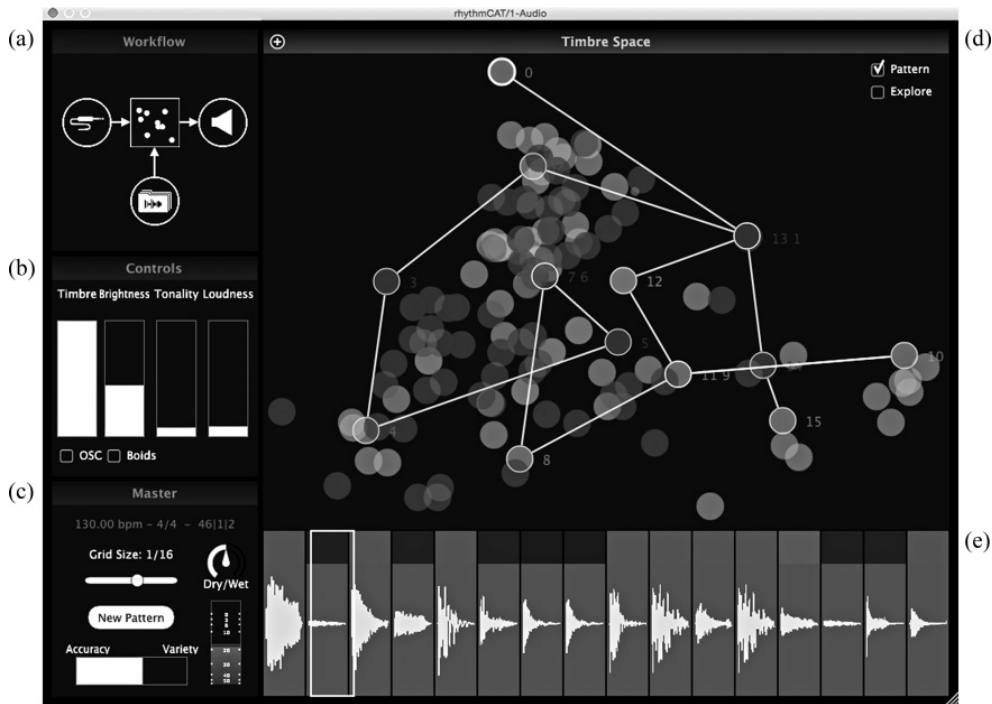
The icons indicate the four main logical operations that the user is likely to implement, and opens up related dialog screens:

- Palette Dialog – indicated by the folder icon
- Seed Dialog – indicated by the jack cable icon
- Sonic Dialog – indicated by the square feature space icon
- Output Dialog – indicated by the speaker icon

#### Sound Palette

The user loads a selection of audio files or folders containing audio files that are analyzed to create the sound palette, as has previously been discussed. Next, dimensionality reduction is performed on each feature vector of the units in the sound palette using principal component analysis (PCA). Two PCA components are retained and scaled to the visible area of the interface to serve as coordinates for placing a circular representation of the sound in two-dimensional space. We call these visual representations, along with their associated audio content, *sound objects*. They are clearly visible in the main Timbre Space window, Figure 2d.

Figure 2. The main user interface for RhythmCat consists of panels for workflow (a), slider controls (b), master controls (c), the main timbre space interface (d), and waveform representation (e).



### Seed Input

Seed audio is captured and analyzed by directly recording the input audio of the track on which the instrument resides in the audio host. Using the real-time tempo and information about bar and beat position provided by the host, the recorder will wait until the next measure starts to begin capture and will only capture complete measures of audio. This audio is analyzed as before, with one exception. Because the goal of the instrument is to integrate with an existing session and generate looped material, we assume that the incoming audio is quantized and matches the tempo of the session. Thus, onset detection is not performed on the seed input; instead, segmentation takes place at the points in time determined by the grid size (lower left of the screen).

An important aspect to note: Because the instrument fundamentally operates in real time, we need

to be careful about performing potentially time-consuming operations, such as feature extraction, when the audio system is running. Thus, we perform the audio-recording stage and feature-extraction process on separate threads, so the main audio-playback thread is uninterrupted. This is separate to yet another thread that handles elements of the user interface.

### Sonic Parameters

Clicking on the square sonic icon in the center of the workflow component opens up the set of sliders shown in Figure 2b, which allows us to adjust the weights of the features in the system. Adjusting these weights has effects in terms of the pattern-generation process but also in the visualization. Presenting their technical names (centroid, flatness, and MFCCs) would be confusing

Figure 3. Algorithm for generating a list of sound connections.

for the general user, so we relabeled them with what we considered the most descriptive subjective terms. With the pattern-generation process, these weights directly affect the features when performing similarity computation and unit selection, as we will see in the next section. Depending on the source and target material, different combinations of feature weightings produce noticeably different results. Informally, we have experienced good results using MFCCs alone, for example, as well as combinations of the flatness and centroid. In terms of visualization, when the weights are changed, dimensionality reduction is reinitiated and, hence, positioning of the sound objects in the timbre space changes. Manipulating these parameters can help disperse and rearrange the sound objects for clearer interaction and exploration by the user in addition to affecting the pattern generation process.

Once the palette and seed matrices have been populated, a similarity matrix between the palette and seed matrix is created. Using the feature weightings from the parameter sliders, a sorted matrix of weighted Euclidean distances between each onset in the target matrix and each unit sound in the palette matrix is computed.

#### Unit Selection and Pattern Generation

The algorithm for unit selection is quite straightforward. For each unit  $i$  in the segmented target sequence (e.g., a 16-step sequence) and each corpus unit  $j$  (typically many more), the target unit cost  $C_{i,j}$  is calculated by the weighted Euclidean distance of each feature  $k$ .

These unit costs are stored in similarity matrix  $M$ . Next we create a matrix  $M'$  of the indices of the elements of  $M$  sorted in ascending order. Finally, a concatenated sequence can be generated by returning a vector of indices  $I$  from this sorted matrix and playing back the associated sound file. To retrieve the closest sequence  $V_0$  one would only need to return the first row.

Returning sequence vectors as rows of a sorted matrix limits the number of possible sequences to the matrix size. This can be extended if we define a similarity threshold  $T$  and return a random index

```

Procedure GET-ONSET-LIST
  for n in GridSize do
    R = Random number 0 < Variance
    I = Index from Row R of Similarity
      Matrix
    S = New SoundConnection
    S->SoundUnit = SoundUnit(I)
    Add S to LinkedList
  end for
return LinkedList
End Procedure

```

between  $0$  and  $j - T$  for each step  $i$  in the new sequence.

When the user presses the New Pattern button (Figure 2c), a new linked list of objects, called *sound connections*, is formed. This represents a traversal through connected sound objects in the timbre space. The length of the linked list is determined by the grid size specified by the user, so if the user specifies, for example, a grid size of 1/16, a one-measure sequence of 16th notes will be generated. The algorithm in Figure 3 details the exact procedure whereby we generate the list. The variance parameter affects the threshold of similarity by which onsets are chosen. With 0 variance, the most similar sequence is always returned. This variance parameter is adjustable from the Accuracy/Variety slider in the lower-left corner of the instrument (Figure 2c).

In the main timbre space interface (Figure 2d), a visual graph is generated in the timbre space by traversing the linked list and drawing line edges connecting each sound object pointed to by the sound connection in the linked list. In this case, a loop of 16 onsets has been generated, with the onset numbers indicated beside the associated sound object for each onset in the sequence. The user is free to manipulate these sound connections to mutate these patterns by touching or clicking on the sound connection and dragging to another sound object. Multiple sound connections assigned to an individual sound object can be selected as a group by slowly double-tapping and then dragging.

On the audio side, every time there is a new beat, the linked list is traversed. If a sound connection's onset number matches the current beat, the corresponding sound unit is played back. One addition

---

that occurred after some user experiments with the prototype is the linear waveform representation of the newly generated sequence (Figure 2e). Users felt the combination of the 2-D interface with the traditional waveform representation made the sequences easier to navigate and they also welcomed being able to manipulate the internal arrangement of sequence itself once generated.

## Evaluation

In the course of our literature review of the state of the art, we were particularly interested in examining the procedures and frameworks used in performing evaluations of the implemented systems. Our most immediate observation was that evaluation is an understudied aspect of research into concatenative systems. With creative and generative systems, this is often the case; many such systems are designed solely with the author as composer in mind.

Some authors provide examples of use cases (Cardle, Brooks, and Robinson 2003). Authors, such as Sturm, have made multimedia examples available on the Web (see Zils and Pachet 2001; Xiang 2002; Sturm 2004). Frequently, researchers have made allusions to some concept of the “user,” but only one paper has presented details of a user experiment (Aucouturier and Pachet 2005). One researcher, Graham Coleman, also highlighted this lack of evaluation strategies in concatenative synthesis in his doctoral dissertation (Coleman 2015). For the evaluation of his own system, he undertook a listening experiment with human participants in tandem with a thorough analysis of algorithmic performance and complexity.

We conducted extensive evaluation of our own system, both quantitatively and qualitatively. In the quantitative portion, we set out to investigate two key aspects. First, if we consider the system as a retrieval task that aims to return similar items, how accurate and predictable is the algorithm and its associated distance metric? Second, how does this objective retrieval accuracy correspond to the perceptual response of the human listener to the retrieved items?

The qualitative evaluation consisted of interactive, informal interviews with intended users—mostly active music producers but also music researchers and students—as they used the software. We gathered their responses and impressions and grouped them according to thematic analysis techniques. As alluded to in the introduction, both the quantitative evaluation and the qualitative evaluation have been previously reported in separate publications, but we include summaries of each here for reference.

## System Evaluation

We describe here the qualitative portion of the evaluation, first by introducing the experimental setup, then presenting and comparing the results of the algorithm’s retrieval accuracy with the listener survey.

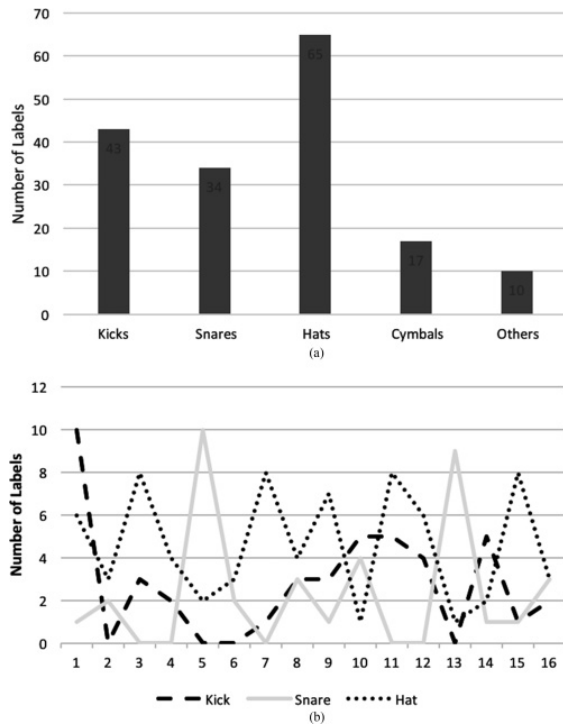
### *Experimental Setup*

Because the goal of the system is the generation of rhythmic loops, we decided to formulate an experiment using breakbeats (short drum solos taken from commercial funk and soul stereo recordings). Ten breakbeats were chosen in the range 75–142 bpm, and we truncated each of them to a single bar in length. Repeating ten times for each loop, we selected a single loop as the target seed and resynthesized it using the other nine loops (similar to holdout validation in machine learning) at four different distances from target to create 40 variations.

Each of the loops was manually labeled with the constituent drum sounds as we hear them. The labeling used was “K” for kick drum, “S” for snare, “HH” for hi-hat, “C” for cymbal, and “X” when the content was not clear (such as artifacts from the onset-detection process or some spillage from other sources in the recording). Figure 4 shows the distribution labels in the entire data set and the distribution according to step sequence. We can notice immediately the heavy predominance of hi-hat sounds, which is typical in kit-based drumming patterns. In addition, the natural trends



Figure 4. Distribution of sound labels in the source corpus (a). Distribution of sound labels by step number in the 16-step sequence (b).



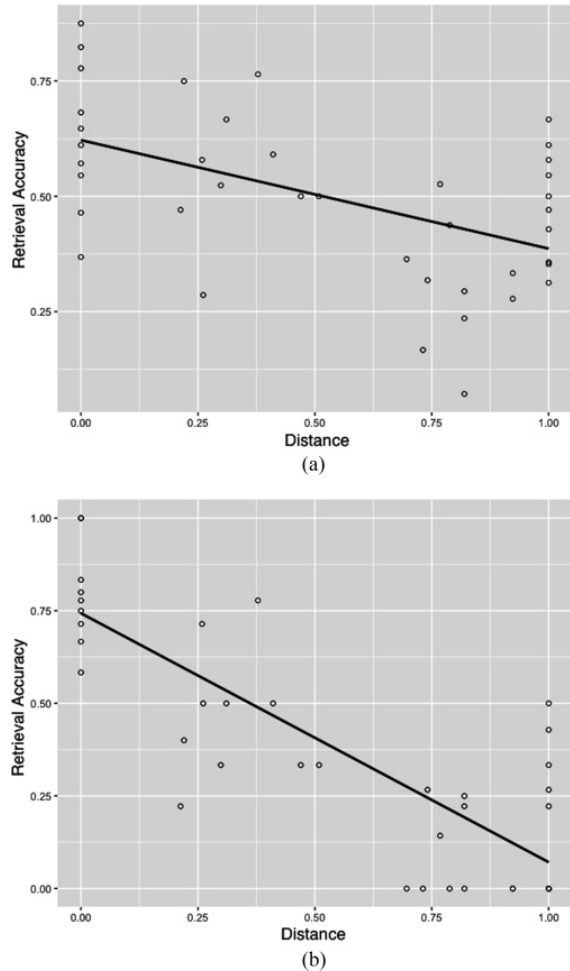
of kit drumming are evident, namely, kick-drum placement on the first beat and offbeat peaks for the snares.

#### Retrieval Evaluation

We compared each of the labels in each 16-step position of the quantized target loop with the labels in each step of the newly generated sequences. The accuracy  $A$  of the algorithm is then given by the number of correctly retrieved labels divided by the total number of labels in the target loop, inspired by a similar approach adopted by Thompson, Dixon, and Mauch (2014).

Based on Pearson's correlation of the retrieval ratings and the distances of the generated patterns, we were able to confirm the tendency of smaller distances to produce more similar patterns in

Figure 5. Scatter plot and linear regression of accuracy versus distance for all sound labels (a) and for the same sequence with kick drum and snare isolated (b).



terms of the labeling accuracy. A moderate negative correlation of  $r = -0.516$  (significance level  $p < 0.001$ ) is visible by the regression line in Figure 5a. If we isolate the kick and snare (often considered the more salient events in drum performances, see Gouyon, Pachet, and Delerue 2000) the negative correlation value decreases sharply to  $r = -0.826$ , as shown in Figure 5b.

---

## Listener Evaluation

Observing that the algorithm tends to reproduce labels in a predictable fashion, we sought to establish whether this conforms in reality to what a human listener perceives. An online listener survey was conducted using the same generated loops and targets from the retrieval evaluation. Twenty-one participants completed the survey, drawn mostly from music researchers and students from the institutions of Universitat Pompeu Fabra and the Escola Superior de Música de Catalunya in Barcelona, as well as friends with an interest in music. Twenty out of those indicated that they played an instrument, with nine specifying an instrument from the percussion family.

The participants were requested to audition a target loop and each of the generated loops in succession. They were then asked to rate, on a Likert scale of 1 to 5, the similarity of the generated loop to the target in terms of their timbre (i.e., do the kick drums, snares, and hi-hats sound alike?) as well as the rhythmic structure of the pattern (i.e., is the arrangement and placement of the sounds similar?). We also asked them to rate their aesthetic preference for the generated patterns, to determine any possible correlation with similarity.

Survey results were collated and analyzed using Spearman's rank correlation, comparing the mode of the participants' responses with the distance value of each loop. A moderate-to-strong negative correlation pattern emerged for all of the variables under consideration, namely,  $r = -0.66$  for pattern similarity,  $r = -0.59$  for timbral similarity, and  $r = -0.63$  for their personal preference according to similarity (with significance levels of  $p < 0.01$  in all instances). It should be evident that the listeners' judgments reflect what the results unearthed in the retrieval evaluation.

## User Evaluation

The quantitative evaluation demonstrated the predictive performance of the algorithm based on retrieval accuracy and the corresponding listeners' judgments of similarity and likeness. Equally deserv-

ing of evaluative scrutiny is the users' experience of working with the software: gauging their responses to the interface, its modes of interactions, and its relevance and suitability for their own compositional styles and processes.

To this effect, a qualitative evaluation phase was arranged to gather rich descriptive impressions from related groups of users in Barcelona and Berlin during February 2016. In Barcelona, as with user profiles of the listener survey, most of the participants were researchers or students in the broad area of Sound and Music Computing. In Berlin we were able to gain access to artists involved in the Red Bull Music Academy as well as with employees of the music software company Native Instruments.

In broad terms, the overall sense of people's impressions was positive. Many participants were initially attracted to the visual nature of the software and were curious to discover its function and purpose. After some familiarization with its operation, people also remarked positively on its sonic output and ability to replicate the target loop:

"It's an excellent tool for making small changes in real time. The interface for me is excellent. This two-dimensional arrangement of the different sounds and its situation by familiarity, it's also really good for making these changes."

"I'm really interested in more-visual, more-graphical interfaces. Also, the fact that you can come up with new patterns just by the push of a button is always great."

"It's inspiring because this mix makes something interesting still, but also I have the feeling I can steal it."

"The unbelievable thing is that it can create something that is so accurate. I wouldn't believe that it's capable of doing such a thing."

Some of the negative criticism came from the prototypical nature of the instrument, and some users were not comfortable with its perceived indeterminacy:

---

"It was too intense and also [had] a prototype feeling. So I was like, 'Well, it's cool and very interesting but not usable yet.'"

"Right now it's still hard to find your way around, but that's something you can refine pretty easily."

#### *Usage Scenarios*

Participants were asked to consider how they would envisage themselves using the software. Most of them concurred that its strength would be in supporting musicians in their production and compositional workflows. Some users were curious about using it in live contexts, such as continuous analysis of instrumental performance or beat-boxing assistance:

"This is great! Ah, but wait . . . Does it mean I could, like, beat box really badly some idea that I have . . . and then bring my samples, my favorite kits and then it will just work?"

Continuous recording and analysis is within the realm of possibility, but can potentially be an operation that is prohibitively computationally expensive, depending on the granularity of the beat grid and the size of the corpus. Further benchmarking and tests are required to establish the upper bounds.

Another interesting observation was that many users did not want to start with a target, preferring to use the instrument as a new, systematic method of exploring their existing sounds:

"I've got this fully on wet straight away, which tells you the direction I'd be going with it."

"You just want to drag in a hundred different songs and you just want to explore without having this connection to the original group. Just want to explore and create sound with it."

#### *Traditional Forms of Navigation*

Our original intention was for users to solely be able to arrange their patterns through the 2-D timbre space. Through the course of our discussions with users we learned that, although they were eager to

adapt the new visual paradigm, they still felt the need for a linear waveform to aid their comprehension. Because of this feedback, the waveform view was implemented early on in our development, as is evident in its inclusion in Figure 2.

"It's a bit hard to figure out which sixteenth you are looking for, because you are so used to seeing this as a step grid."

"You have a waveform or something. . . Then I know, okay, this is the position that I'm at."

"Is there also a waveform place to put the visualization? People are so used to having that kind of thing."

#### *Shaping Sounds*

A recurring issue, which cropped up mainly with producers and DJs, was the desire to shape, process, and refine the sounds once a desirable sequence was generated by the system. This way of composing seems emblematic of electronic music producers across the board; they start with a small loop or idea then vary and develop it exploiting the many effects processing and editing features provided by their tools. Most crucially, they desired the option to be able to control the envelopes of the individual units via drawable attack and decay parameters, which is currently being implemented.

" . . . an attack and decay just to sort of tighten it up a little bit . . . get rid of some of the rough edges of the onsets and offsets."

"It would be great if you could increase the decay of the snare, for example. Which, if it's prototype, you can't expect to have all those functions there immediately, but in an end product, I think it would be a necessity."

#### *Parameterization and Visualization*

The most overarching source of negative criticism from all users was in how we presented the parameters of the system. Users are freely able to manipulate the individual weightings of the features, affecting their relative influence in

---

the similarity computation, but also in the PCA dimensional-reduction stage. In an effort to make this more “user friendly,” we relabeled the feature names with more generally comprehensible terms like “timbre,” “brightness,” “harmonicity,” and “loudness.” Despite this, participants reported being confused and overwhelmed by this level of control, stating that they were “a bit lost already,” that there are “four parameters, and you don’t know which thing is what,” and that they “would prefer not to have too many controls.”

Most users were quite content with the overall sonic output from the system without delving into the manipulation of feature parameters. For the visualization, however, there are certain configurations of the features that produce the best separation and clustering of the units (although MFCCs alone appear to be the most robust in our experience).

One option we are actively investigating would be to remove these parameter sliders and replace them with an optional “advanced” mode, giving users the ability to select specific parameters for the axes (as in CataRT) in addition to “automatic” arrangement configurations made possible by using dimensionality-reduction techniques. These configurations could be derived by analyzing different sound sets to find weighting combinations that give the best visual separation, depending on the corpus provided. Finally, we are currently using PCA for dimensionality reduction. There are also other approaches, including multidimensional scaling (Donaldson, Knopke, and Raphael 2007) and the recent t-distributed stochastic neighbor embedding algorithm (Erisson 2015; Turquois et al. 2016), which have been used in musically related tasks and that we are implementing and evaluating as alternatives.

## Discussion

Evaluating systems for music creation and manipulation is a difficult, ill-defined, and insufficiently reported task. As we have stressed in the course of this article, this is also the case with systems for concatenative synthesis. After conducting our own evaluation, we considered what key points could be made to help inform future evaluations by interested

researchers in the community. Our observations led us to indicate three distinct layers that should be addressed for a significant, full-fledged appraisal.

The most high-level and general “system” layer calls for user evaluations that go beyond “quality of experience” and “satisfaction” surveys. Such evaluations should strive to address creative productivity and workflow efficiency aspects particular to the needs of computer-music practitioners.

At the mid-level “algorithmic” layer, we examine the mechanics of developing solutions strategies for concatenative synthesis. We have identified three main trends in algorithmic techniques used for tackling tasks in concatenative synthesis, namely, similarity-matrix and clustering approaches (like ours), Markov models, and constraint-satisfaction problems. Each of these techniques exhibits its own strengths and weaknesses in terms of accuracy, flexibility, efficiency, and complexity. Comparing these algorithms within a single system and, indeed, across multiple systems, using a well-defined data set, a clear set of goals, and specific success criteria would represent a valuable asset in the evaluation methodology of concatenative synthesis. Additionally, we should pay attention to the distance and similarity metrics used, as there are other possibilities that are explored and compared in other retrieval problems (e.g., Charulatha, Rodrigues, and Chitralkha 2013).

At the lowest level, the focus is on the broader implications related to MIR of choosing appropriate features for the task at hand. In the course of our evaluation, we chose the features indicated in the implementation and did not manipulate them in the experiment. There are, of course, many other features relevant to the problem that can be studied and estimated in a systematic way, as is par for the course in classification experiments in MIR. Furthermore, tuning the weights was not explored and is an important consideration that depends greatly on different corpora and output-sequence requirements.

In addition to this three-tiered evaluation methodology, an ideal component would be the availability of a baseline or comparison system that ensures new prototypes improve over some clearly identifiable aspect. Self-referential evaluations run the risk of

---

confirming experimenter bias without establishing comprehensive criticism.

## Conclusion

In this article, we explored concatenative synthesis as a compositional tool for generating rhythmic patterns for electronic music, with a strong emphasis on its role in EDM musical styles. One of our first contributions was to present a thorough and up-to-date review of the state of the art, beginning with its fundamental algorithmic underpinnings and proceeding to modern systems that exploit new and experimental visual and interactive modalities. Although there are a number of commercial applications that encapsulate techniques of concatenative synthesis for the user, the vast majority of systems are frequently custom-built for the designer or are highly prototypical in nature. Consequently, there is a marked lack of evaluation strategies or reports of user experiences in the accompanying literature.

Based on these investigations, we set out to design a system that applied and extended many of the pervasive techniques in concatenative synthesis with a clear idea of its application and its target user. We built an instrument that was easily integrated with modern digital audio workstations and presented an interface that intended to be attractive and easy to familiarize oneself with. How to evaluate the system, not only in terms of its objective performance but also in its subjective aural and experiential implications for our users, was our final substantial contribution to this area. The results of our evaluations showed that our system performed as expected, and users were positive about its potential for assisting in their creative tasks, while also proposing interesting avenues for future work and contributions.

## Resources

A demonstration version of the software is available online at <http://github.com/carthach/rhythmCAT>. A video example can be viewed at <http://youtu.be/hByhgFfzto>.

## References

- Aucouturier, J.-J., and F. Pachet. 2005. "Ringomatic: A Real-Time Interactive Drummer Using Constraint-Satisfaction and Drum Sound Descriptors." *Proceedings of the International Conference on Music Information Retrieval*, pp. 412–419.
- Bartók, B. 1993. "Hungarian Folk Music." In B. Suchoff, ed. *Béla Bartók Essays*. Lincoln: University of Nebraska Press, pp. 3–4.
- Battier, M. 2007. "What the GRM Brought to Music: From *Musique Concrète* to Acousmatic Music." *Organized Sound* 12(3):189–202.
- Bello, J., and L. Daudet. 2005. "A Tutorial on Onset Detection in Music Signals." *IEEE Transactions on Audio, Speech, and Language Processing* 13(5):1035–1047.
- Bernardes, G., C. Guedes, and B. Pennycook. 2013. "EarGram?: An Application for Interactive Exploration of Concatenative Sound Synthesis in Pure Data." In M. Aramaki et al., eds. *From Sounds to Music and Emotions*. Berlin: Springer, pp. 110–129.
- Böck, S., and G. Widmer. 2013. "Maximum Filter Vibrato Suppression for Onset Detection." In *Proceedings of the International Conference on Digital Audio Effects*. Available online at [dafx13.nuim.ie/papers/09.dafx2013.submission.12.pdf](http://dafx13.nuim.ie/papers/09.dafx2013.submission.12.pdf). Accessed January 2017.
- Bogdanov, D., et al. 2013. "ESSENTIA: An Audio Analysis Library for Music Information Retrieval." In *Proceedings of the International Conference on Music Information Retrieval*, pp. 493–498.
- Bradski, G. 2000. "The OpenCV Library." *Dr. Dobb's Journal* 25(11):120–125.
- Brent, W. 2010. "A Timbre Analysis and Classification Toolkit for Pure Data." In *Proceedings of the International Computer Music Conference*, pp. 224–229.
- Cardle, M., S. Brooks, and P. Robinson. 2003. "Audio and User Directed Sound Synthesis." *Proceedings of the International Computer Music Conference*, pp. 243–246.
- Charulatha, B., P. Rodrigues, and T. Chitrakleha. 2013. "A Comparative Study of Different Distance Metrics That Can Be Used in Fuzzy Clustering Algorithms." *International Journal of Emerging Trends and Technology in Computer Science*. Available online at [www.ijetcs.org/NCASG-2013/NCASG\\_38.pdf](http://www.ijetcs.org/NCASG-2013/NCASG_38.pdf). Accessed January 2017.
- Coleman, G. 2015. "Descriptor Control of Sound Transformations and Mosaicing Synthesis." PhD dissertation,

- Universitat Pompeu Fabra, Department of Information and Communication Technologies, Barcelona.
- Dixon, S. 2006. "Onset Detection Revisited." In *Proceedings of the International Conference on Digital Audio Effects*, pp. 133–137.
- Donaldson, J., I. Knopke, and C. Raphael. 2007. "Chroma Palette?: Chromatic Maps of Sound as Granular Synthesis Interface." In *Proceedings of the Conference on New Interfaces for Musical Expression*, pp. 213–219.
- Frisson, C. 2015. "Designing Interaction for Browsing Media Collections (by Similarity)." PhD dissertation, Université de Mons, Faculty of Engineering.
- Frisson, C., C. Picard, and D. Tardieu. 2010. "Audiogarden?: Towards a Usable Tool for Composite Audio Creation." *QPSR of the Numediart Research Program* 3(2):33–36.
- Gouyon, F., F. Pachet, and O. Delerue. 2000. "On the Use of Zero-Crossing Rate for an Application of Classification of Percussive Sounds." In *Proceedings of the International Conference on Digital Audio Effects*, pp. 3–8.
- Greenberg, C. 1971. "Collage." In *Art and Culture: Critical Essays*. Boston, Massachusetts: Beacon Press, pp. 70–83.
- Hackbarth, B., N. Schnell, and D. Schwarz. 2011. "AudioGuide?: A Framework for Creative Exploration of Concatenative Sound Synthesis." IRCAM Research Report. Available online at [articles.ircam.fr/textes/Hackbarth10a/index.pdf](http://articles.ircam.fr/textes/Hackbarth10a/index.pdf). Accessed January 2017.
- Herrera, P., A. Dehamel, and F. Gouyon. 2003. "Automatic Labeling of Unpitched Percussion Sounds." In *Proceedings of the 114th Audio Engineering Society Convention*. Available online at [www.aes.org/e-lib/browse.cfm?elib=12599](http://www.aes.org/e-lib/browse.cfm?elib=12599) (subscription required). Accessed January 2017.
- Holmes, T. 2008. *Electronic and Experimental Music*. Abingdon-on-Thames, UK: Routledge.
- Hoskinson, R., and D. Pai. 2001. "Manipulation and Resynthesis with Natural Grains." In *Proceedings of the International Computer Music Conference*, pp. 338–341.
- Hoskinson, R., and D. K. Pai. 2007. "Synthetic Soundscapes with Natural Grains." *Presence: Teleoperators and Virtual Environments* 16(1):84–99.
- Humphrey, E. J., D. Turnbull, and T. Collins. 2013. "A Brief Review of Creative MIR." In *Proceedings of the International Conference on Music Information Retrieval*. Available online at [ismir2013.ismir.net/wp-content/uploads/2014/02/lbd1.pdf](http://ismir2013.ismir.net/wp-content/uploads/2014/02/lbd1.pdf). Accessed January 2017.
- Hunt, A. J., and A. W. Black. 1996. "Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database." In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 373–376.
- Jehan, T. 2005. "Creating Music by Listening." PhD dissertation, Massachusetts Institute of Technology, Media Arts and Sciences.
- Knees, P., et al. 2016. "The GiantSteps Project: A Second-Year Intermediate Report." In *Proceedings of the International Computer Music Conference*, pp. 363–368.
- Lochhead, J. I., and J. H. Auner. 2002. *Postmodern Music/Postmodern Thought: Studies in Contemporary Music and Culture*. Abingdon-on-Thames, UK: Routledge.
- Logan, B. 2000. "Mel Frequency Cepstral Coefficients for Music Modeling." In *Proceedings of the International Symposium on Music Information Retrieval*. Available online at [ismir2000.ismir.net/papers/logan\\_paper.pdf](http://ismir2000.ismir.net/papers/logan_paper.pdf). Accessed January 2017.
- Loughran, R., et al. 2004. "The Use of Mel-Frequency Cepstral Coefficients in Musical Instrument Identification." In *Proceedings of the International Computer Music Conference*, pp. 42–43.
- Mangani, M., R. Baldizzone, and G. Nobile. 2006. "Quotation in Jazz Improvisation: A Database and Some Examples." Paper presented at the International Conference on Music Perception and Cognition, August 22–26, Bologna, Italy.
- Masri, P., and A. Bateman. 1996. "Improved Modeling of Attack Transients in Music Analysis-Resynthesis." In *Proceedings of the International Computer Music Conference*, pp. 100–103.
- McLeod, K. 2009. "Crashing the Spectacle: A Forgotten History of Digital Sampling, Infringement, Copyright Liberation and the End of Recorded Music." *Culture Machine* 10:114–130.
- Metzer, D. 2003. *Quotation and Cultural Meaning in Twentieth-Century Music*. Cambridge: Cambridge University Press.
- Miller, P. D. 2008. *Sound Unbound: Sampling Digital Music and Culture*. Cambridge, Massachusetts: MIT Press.
- O'Connell, J. 2011. "Musical Mosaicing with High Level Descriptors." Master's thesis, Universitat Pompeu Fabra, Sound and Music Computing, Barcelona.
- Ó Nuanáin, C., P. Herrera, and S. Jordà. 2016. "An Evaluation Framework and Case Study for Rhythmic Concatenative Synthesis." In *Proceedings of the International Society for Music Information Retrieval Conference*, pp. 67–72.

- 
- Ó Nuanáin, C., S. Jordà, and P. Herrera. 2016a. "An Interactive Software Instrument for Real-Time Rhythmic Concatenative Synthesis." In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pp. 383–387.
- Ó Nuanáin, C., S. Jordà, and P. Herrera. 2016b. "Towards User-Tailored Creative Applications of Concatenative Synthesis in Electronic Dance Music." In *Proceedings of the International Workshop on Musical Metacreation*. Available online at [musicalmetacreation.org/buddydrive/file/nuanain\\_towards](http://musicalmetacreation.org/buddydrive/file/nuanain_towards). Accessed January 2017.
- Oswald, J. 1985. "Plunderphonics, or Audio Piracy as a Compositional Prerogative." Paper presented at the Wired Society Electro-Acoustic Conference, Toronto, Canada. Reprinted in *Musicworks*, Winter 1986, 34:5–8.
- Peeters, G. 2004. "A Large Set of Audio Features for Sound Description (Similarity and Classification) in the CUIDADO Project." IRCAM Project Report. Available online at [recherche.ircam.fr/anasyn/peeters/ARTICLES/Peeters\\_2003\\_cuidadoaudiofeatures.pdf](http://recherche.ircam.fr/anasyn/peeters/ARTICLES/Peeters_2003_cuidadoaudiofeatures.pdf). Accessed January 2017.
- Rabiner, L. R. 1989. "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition." *Proceedings of the IEEE* 77(2):257–286.
- Roads, C. 2004. *Microsound*. Cambridge, Massachusetts: MIT Press.
- Rodgers, T. 2003. "On the Process and Aesthetics of Sampling in Electronic Music Production." *Organized Sound* 8(3):313–320.
- Roy, P., F. Pachet, and S. Krakowski. 2007. "Analytical Features for the Classification of Percussive Sounds: The Case of the Pandeiro." In *Proceedings of the International Conference on Digital Audio Effects*, pp. 213–220.
- Russell, S., and P. Norvig. 2009. *Artificial Intelligence: A Modern Approach*. Upper Saddle River, New Jersey: Prentice Hall.
- Schwarz, D. 2003. "The Caterpillar System for Data-Driven Concatenative Sound Synthesis." In *Proceedings of the International Conference on Digital Audio Effects*, pp. 135–140.
- Schwarz, D. 2005. "Current Research in Concatenative Sound Synthesis." In *Proceedings of the International Computer Music Conference*, pp. 9–12.
- Schwarz, D. 2017. "Corpus-Based Sound Synthesis Survey." Available online at [imtr.ircam.fr/imtr/Corpus-Based\\_Sound\\_Synthesis\\_Survey](http://imtr.ircam.fr/imtr/Corpus-Based_Sound_Synthesis_Survey). Accessed February 2017.
- Schwarz, D., et al. 2006. "Real-Time Corpus-Based Concatenative Synthesis with CataRT." In *Proceedings of the International Conference on Digital Audio Effects*, pp. 18–21.
- Sewell, A. 2014. "Paul's Boutique and Fear of a Black Planet: Digital Sampling and Musical Style in Hip Hop." *Journal of the Society for American Music* 8(1):28–48.
- Sturm, B. L. 2004. "Matconcat: An Application for Exploring Concatenative Sound Synthesis Using Matlab." In *Proceedings of the International Conference on Digital Audio Effects*, pp. 323–326.
- Sturm, B. L. 2006. "Adaptive Concatenative Sound Synthesis and Its Application to Micromontage Composition." *Computer Music Journal* 30(4):46–66.
- Tamagawa, K. 1988. "Echoes from the East: The Javanese Gamelan and Its Influence on the Music of Claude Debussy." DMA dissertation, University of Texas at Austin.
- Thompson, L., S. Dixon, and M. Mauch. 2014. "Drum Transcription via Classification of Bar-Level Rhythmic Patterns." In *Proceedings of the International Society for Music Information Retrieval Conference*, pp. 187–192.
- Tindale, A., et al. 2004. "Retrieval of Percussion Gestures Using Timbre Classification Techniques." *Proceedings of the International Conference on Music Information Retrieval*, pp. 541–545.
- Turquois, C., et al. 2016. "Exploring the Benefits of 2D Visualizations for Drum Samples Retrieval." In *Proceedings of the ACM SIGIR Conference on Human Information Interaction and Retrieval*, pp. 329–332.
- Tzanetakis, G., G. Essl, and P. Cook. 2001. "Automatic Musical Genre Classification of Audio Signals." In *Proceedings of the International Symposium on Music Information Retrieval*, pp. 293–302.
- Xenakis, I. 1971. *Formalized Music*. Bloomington: Indiana University Press.
- Xiang, P. 2002. "A New Scheme for Real-Time Loop Music Production Based on Granular Similarity and Probability Control." In *Proceedings of the International Conference on Digital Audio Effects*, pp. 89–92.
- Zils, A., and F. Pachet. 2001. "Musical Mosaicing." In *Proceedings of the International Conference on Digital Audio Effects*, 1–6. Available online at [www.csis.ul.ie/dafx01/proceedings/papers/zils.pdf](http://www.csis.ul.ie/dafx01/proceedings/papers/zils.pdf). Accessed January 2017.





## 6. CONCLUDING THOUGHTS

*You never know what is enough  
unless you know what is more than enough*

William Blake  
*The Marriage of Heaven and Hell*

I started my journey in MIR with some ideas and targets that were motivated by my own experience in music reinforcement, mixing and creation. I “wanted” to see realized some of my dreams, and the promise of technologies to lead me there, kept me challenged and interested in the field as an active researcher. Twenty years later the dream is half reality and half nightmare. As researchers, we are partially responsible of the growing banalization of music as an experience (I’m not talking about content here). Having helped to get “whatever music whenever and wherever you are” we have contributed to diminish the value given to concerts, shared attentive listening sessions, extremely moving personal experiences, or paying for specific multisensorial events. Sometimes, technologies that were probably developed with a deep love and enthusiasm for music, have turned into weapons against its human inceptors. Sometimes I have the impression that we might had been playing “sorcerer’s apprentice” without being aware of that, generating knowledge (for free) with which big companies made the money and left us with just a ten percent of fun and functionalities, considering what could be done with our findings and creations. Not to mention how biased by our own cultural values and limited musical experience our thoughts and algorithms embedding those thoughts have been (and still are). And, to conclude, how far of real people much of our research has been done? I think our research community has been too solipsist, tending to consider (as a justification mechanism) that everybody is so music nerd as we are. And that is not true. We want everybody to be engaged in music activities everywhere, all the time, we thought that many people were waiting for humming queries all day, or to navigating across iconified music collections as a nicer view than their daily navigation from home to work, or that they would be eager to music discoveries, looking for at least one novel music item a day... but then we, both as collectives and individually, rarely attend to the concerts we think are worth, or give support (financial, I mean) to the artists we think they deserve it, or do not give quality time to listen to “that” nice piece of music we discovered yesterday.

This thesis, though, is not about sociology of technology and my criticism is probably linked to the fact that I have got old and grumpy, and that my memory leaks more than what I would like. I tried here to organize some snapshots of my journey and achievements in a coherent and meaningful way, by means of the idea of ages that start when certain shortcomings are evident and, at the same time, candidate technologies for overcoming them are available. From the age of feature extractors to the age of creative systems, I have presented selected papers of research that were framed in them. Even though my deepest and probably fruitful involvement happened in the age of semantic

descriptors, I still look forward the promises of mixing MIR and HCI in the age of creative systems.

Before 1998 there was just a single conference to present research and development on music and technology, the ICMC (International Computer Music Conference). Twenty years later you can find every three months a workshop or conference that, even in the case of not being focused on musical issues, will accept and properly review papers on that subject. ISMIR is the conference with the highest impact factor in the fields of music and musicology, according to Google Scholar<sup>46</sup>, but there are some others which are highly influential such as NIME (New Interfaces for Musical Expression), DAFx (Digital Audio Effects) or SMC (Sound and Music Computing), which also accept papers on MIR (even the ICASSP, the fourth international conference in signal processing<sup>47</sup>, has included sessions on MIR in recent years). This is another proof of the good health of our field. Looking at how the technologies have evolved, it is like comparing night and day. Polyphonic music can be described (even unmixed) with a level of detail and quality that is making possible professional applications. Recent papers, thanks to the intervention of deep learning, report on improvements (from promising to substantial) on every musical facet that has traditionally been addressed in MIR (from tempo estimation to mood classification, from tonal to structural description) (Choi et al., 2017c; Müller et al., 2017). Music recommendation is reliable and relevant, at least, when it comes from the main providers. Even though all that bright scenario, when I finish reading many papers I, nevertheless, still wonder with a deadpan face: “what did this paper taught me about music”?

Using Kuhn’s terminology (2012) it seems as MIR is reaching the “paradigmatic” status, which means its members share a set of tacit assumptions about the way to approach music research (that we need features, that we need computable concepts, that algorithms have to be evaluated, that context is essential...), a research agenda (which are the essential topics?) and recipes to address those different problems. This status facilitates that new practitioners start working in the field in a sound and safe way, and that they beforehand know about what peers are expecting when reviewing their work. This status may be problematic, at the same time, as “exotic” approaches or “new” topics may find resistance to be accepted. In any case, it is refreshing to witness the efforts that some researchers are doing to challenge the approaches, methods and even the relevance of certain topics (Aucouturier & Bigand, 2013; Sturm, 2014; Sturm & Collins, 2014; Sturm et al., 2014; Sturm, 2015). This is, no doubt, a sign of good health.

In this account I have depicted the early scenario when PhD students spent years to discover or invent good features and pre-processing strategies (and then tweaking them) to tackle a music description problem having to do with timbre, or rhythm, or tonality, or genre classification... These days, a bread-and-butter Convolutional Neural Network, or a Recurrent Neural Network combined with Restricted Boltzmann Machines (RNN-RBM) will surpass the old features in one week or less. And the same can be said for whatever music classification problem that can be addressed. But, during the ages of MIR, we learned a lot about music and audio while searching for good features to detect guitar

---

<sup>46</sup> [https://scholar.google.es/citations?view\\_op=top\\_venues&hl=en&vq=hum\\_musicmusicology](https://scholar.google.es/citations?view_op=top_venues&hl=en&vq=hum_musicmusicology)

<sup>47</sup> [https://scholar.google.es/citations?view\\_op=top\\_venues&hl=en&vq=eng\\_signalprocessing](https://scholar.google.es/citations?view_op=top_venues&hl=en&vq=eng_signalprocessing)

solos (Fuhrmann et al., 2009), mellotrons (Román-Echeverri & Herrera, 2013), mysterious or humourous music (Sarasúa et al., 2012), infant-oriented singing (Salselas & Herrera, 2011) or to rebuild a personal soundtrack to Alzheimer's patients (Navarro et al., 2014). And all that still counts more, for me, than an accuracy increase of five percent units. Fortunately, we still need to keep human listeners and players close to the lab as their semantic mapping of the world is in constant evolution (that's adaptation, in fact), and the ways music was tagged ten years ago will be different in ten years more, so the semantic path will always be open. I still perceive reticence in some people when they are faced with the idea to computationally capture a portion of their mental models of music, and this means there is still room for improvement. The discoveries made in the neurosciences of music, push us to envision personalisation experiences that decades ago belonged to science fiction. The status of "permanent sensorium" that (thanks to portable devices) the world has become, will give us precious data to understand how our music experiences are shaped by contextual factors of yet unsuspected influence (and hence how can be artificially altered to our convenience... or maybe to the convenience of the technological suppliers?). I hope to live enough to get my hands dirty with some chunks of those data and knowledge, in order to improve our current understanding of music understanding. And, when I retire, I hope to be allowed to play, in the nursing home, with fun music creation systems that can revive my personal memories in original ways, improve my social skills by means of the involvement in an orchestra of tangible wheelchairs, or let me really feel that, instead of being interfaced, I AM the music that I am playing.



## REFERENCES

- Aljanaki A., Yang, Y-H, Soleymani, M. (2017) Developing a benchmark for emotional analysis of music. *PLoS ONE* 12(3): e0173392.
- Andersen, H. K. G., & Knees, P. (2016). *Conversations with expert users in music retrieval and research challenges for creative MIR*. Paper presented at 17th International Society for Music Information Retrieval Conference, 122-128.
- Anderson, L.W. & Krathwohl, D.R. (Eds.) (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. New York: Addison Wesley Longman.
- Arcos, J. L., López de Mántaras, R. & Serra, X. (1998). Generating expressive musical performances with SaxEx, *Journal of New Music Research*, 27(3), 194-210.
- Assayag, G., Bloch, G., Cont, A. & Dubnov, S. (2010). Interaction with machine improvisation. In S. Argamon, K. Burns & S. Dubnov (eds.), *The Structure of Style*. Springer, 219–245.
- Aucouturier, J.-J. & Bigand, E. (2013). Seven problems that keep MIR from attracting the interest of cognition and neuroscience. *Journal of Intelligent Information Systems* 41, 483-497.
- Aucouturier, J.-J., & Pachet, F. (2002). *Music similarity measures: what's the use?* Paper presented at 3rd International Symposium on Music Information Retrieval. Paris, France.
- Aucouturier, J.-J., & Pachet, F. (2004). Improving timbre similarity: How high is the sky? *Journal of Negative Results in Speech and Audio Sciences*, 1(1).
- Bandiera, G. (2015). A content-aware interactive explorer of digital music collections: The Phonos music explorer. MSc Thesis, University of Padova, Padova, Italy.
- Ballet G., Borghesi, R., Hoffmann, P. Lévy, F. (1999). *Studio online 3.0 : An internet "killer application" for remote access to Ircam sounds and processing tools*, Paper presented at Journées d'Informatique Musicale, Paris, France.
- Baltrunas, L., & Amatriain, X. (2009). *Towards time-dependant recommendation based on implicit feedback*. Paper presented at RecSys09 Workshop on Context-aware RecommenderSystems (CARS-2009), New York, NY, USA.
- Baltrunas, L., Kaminskas, M., Ludwig, B., Moling, O., Ricci, F., Lüke, K. H. & Schwaiger, R. (2011). *InCarMusic: Context-aware music recommendations in a car*. Paper presented at International Conference on Electronic Commerce and Web Technologies (EC-Web). Toulouse, France.
- Barrington, L., Turnbull, D., Yazdani, M., & Lanckriet, G. (2009). *Combining audio content and social context for semantic music discovery*. Paper presented at SIGIR, Boston, MA, USA, 387-394.
- Baumann, S. & Hummel, O. (2005). Enhancing music recommendation algorithms using cultural metadata. *Journal of New Music Research*, 34(2):161-172.
- Baur, D. (2011). *The songs of our past: Working with listening histories*. Paper presented at ACM Computer Human Interaction 2011 (CHI-2011), Vancouver, BC, Canada.

- Baur, D., Seiffert, F., Sedlmair, M., Boring, S. (2010). The streams of our lives: visualizing listening histories in context. *IEEE Transactions on Visualization and Computer Graphics*, 16(6), 1119-1128.
- Baur, D., Langer, T., & Butz, A. (2009). *Shades of music: Letting users discover sub-song similarities*. Paper presented at 10th International Society for Music Information Retrieval Conference (ISMIR 2009), Kobe, Japan, 111-116.
- Berenzweig, A., Logan, B., Ellis, D. P. W., & Whitman, B. (2004). A large-scale evaluation of acoustic and subjective music similarity measures. *Computer Music Journal*, 28(2), 63-76.
- Bernardi, L., Porta, C., Casucci, G., Balsamo, R., Bernardi, N., Fogari, R. et al. (2009). Dynamic interactions between musical, cardiovascular, and cerebral rhythms in humans. *Circulation*, 119(25), 3171-3180.
- Bernatzky, G., Presch, M., Anderson, M., & Panskepp, J. (2011). Emotional foundations of music as a non-pharmacological pain management tool in modern medicine. *Neuroscience and Biobehavioral Reviews*, 35(9), 1989-1999.
- Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: Handbook I: Cognitive domain*. New York: David McKay.
- Bogdanov, D., Haro, M., Fuhrmann, F., Xambó, A., Gómez, E., & Herrera, P. (2011). *A content-based system for music recommendation and visualization of user preferences working on semantic notions*. Paper presented at 9th IEEE International Workshop on Content-based Multimedia Indexing, Madrid, 249-252.
- Bogdanov, D., Serrà, J., Wack, N., & Herrera, P. (2009). *From low-level to high-level: comparative study of music similarity measures*. Paper presented at IEEE Int. Symp. on Multimedia, Workshop on Advances in Music Information Research (AdMIRe), San Diego, CA, USA, 453-458.
- Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Herrera, P., Mayor, O., Roma, G., Salamon, J., Zapata, J. & Serra, X. (2013a). *ESSENTIA: an audio analysis library for music information retrieval*. Paper presented at International Society for Music Information Retrieval Conference (ISMIR), Curitiba, Brazil. 493-498.
- Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Herrera, P., Mayor, O., Roma, G., Salamon, J., Zapata, J. & Serra, X. (2013b). *ESSENTIA: an open-source library for sound and music analysis*. Paper presented at ACM International Conference on Multimedia (MM'13), Brisbane, Australia, 855-858.
- Bogdanov, D., Wack, N., Gómez, E., Gulati S., Herrera, P., Mayor, O., Roma, G., Salamon, J., Zapata, J. & Serra, X. (2014). "ESSENTIA: an open source library for audio analysis". *ACM SIGMM Records*. 6(1).
- Bonnin, G. & Jannach, D. (2014). Automated generation of music playlists: Survey and experiments. *Comput. Surveys*, 47(2), 1-35.
- Bradley, M. M., & Lang, P. J. (2007). *International affective digitized sounds (2nd Edition; IADS-2): Affective ratings of sounds and instruction manual (Technical Report No. B-3)*. Gainesville, FL: University of Florida, NIMH Center for the Study of Emotion and Attention.

- Bregman, A. S. (1990). Auditory scene analysis. Harvard, MA: Massachusetts Institute Press.
- Brochu, E., De Freitas, N., & Bao, K. (2003). *The sound of an album cover: Probabilistic multimedia and IR*. Paper presented at Workshop on Artificial Intelligence and Statistics, Key West, FL, USA.
- Brodeur M.B., Dionne-Dostie, E., Montreuil, T., Lepage, M. (2010) The bank of standardized stimuli (BOSS), a New Set of 480 Normative Photos of Objects to Be Used as Visual Stimuli in Cognitive Research. *PLOS ONE*, 5, e10773.
- Cabredo, R., Legaspi, R., & Numa, M. (2011). *Identifying emotion segments in music by discovering motifs in physiological data*. Paper presented at 12th International Conference on Music Information Retrieval, Miami, FL, USA, 753-758.
- Cano, P., Koppenberger, M., Herrera, P., Celma, O., & Tarasov, V. (2004a). *Sound effects taxonomy management in production environments*. Paper presented at 25th International AES Conference, London.
- Cano, P., Koppenberger, M., Le Goux, S., Ricard, J., Herrera, P., & Wack, N. (2004b). *Nearest-neighbor generic sound classification with a wordnet-based taxonomy*. Paper presented at 116th Convention of the Audio Engineering Society, Berlin, Germany.
- Cano, P., Koppenberger, M., Le Goux, S., Ricard, J., Herrera, P., & Wack, N. (2005). Nearest-neighbor automatic sound annotation with a Wordnet taxonomy. *Journal of Intelligent Information Systems*, 24(2-3), 99-111.
- Casey, M. A. (1998). Auditory Group Theory: with applications to statistical basis methods for structured audio. PhD thesis. Massachusetts Institute of Technology, Cambridge, MA, USA.
- Casey, M. A. (2002). Sound classification and similarity. In B.S. Manjunath, P. Salembier, & T. Sikora (Eds.), *Introduction to MPEG-7: Multimedia content description language* (pp. 153-164).
- Celma, O., Herrera, P., & Serra, X. (2006). *Bridging the music semantic gap*. Paper presented at Workshop on Mastering the Gap: From Information Extraction to Semantic Representation, 3rd Annual European Semantic Web Conference, Budva, Montenegro.
- Chai, W. (2005). Automated analysis of musical structure. PhD thesis. Massachusetts Institute of Technology, Cambridge, MA, USA.
- Cheng, Z., & Shen, J (2014). *Just-for-Me: an adaptive personalization system for location-aware social music recommendation*. Paper presented at 4th ACM International Conference on Multimedia Retrieval (ICMR). Glasgow, UK.
- Choi, K., Fazekas, G., Sandler, M., & Cho, K. (2017a). *Convolutional recurrent neural networks for music classification*. Paper presented at IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), New Orleans, LA, 2392-2396.
- Choi, K., Fazekas, G., Sandler, M., & Cho, K. (2017b). *Transfer learning for music classification and regression tasks*. Paper presented at 18th International Society of Music Information Retrieval (ISMIR) Conference, Suzhou, China.

- Choi, K., Fazekas, G., Cho, K., & Sandler, M.B. (2017c). *A Tutorial on Deep Learning for Music Information Retrieval*. CoRR, abs/1709.04396.
- Chupchik, G. C., Rickert, M., & Mendelson, J. (1982). Similarity and preference judgements of musical stimuli. *Scandinavian Journal of Psychology*, 23, 273-282.
- Cliff, D. (2000). Hang the DJ: Automatic sequencing and seamless mixing of dance-music tracks. Hp Laboratories Technical Report Hpl, 104:1–11, 2000.
- Collins, N. (2002). *The BBCut Library*. Paper presented at International Computer Music Conference, Goteborg, Sweden.
- Cope, D. (1991). *Computers and musical style*. Oxford, UK: Oxford University Press.
- Davies, M. E. Hamel, P., Yoshii, K. & Goto, M. (2014). AutoMashUpper: Automatic creation of multi-song music mashups. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(12), 1726–1737.
- Davies M. E. & Plumbley, M. D. (2007) Context-dependent beat tracking of musical audio, *IEEE Transactions on Audio, Speech, and Language Processing*, 15 (3), 1009–1020.
- Dey, A. K. (2001). Understanding and using context. *Personal Ubiquitous Computing*, 5(1), 4-7.
- Dey, A. K., & Abowd, G. D. (2000). *Towards a better understanding of context and context-awareness*. Workshop on the What, Who, Where, When, Why and How of Context-Awareness. Paper presented at Computer Human Interaction (CHI-2000), The Hague, The Netherlands.
- Dieleman, S. & Schrauwen, B. (2014). *End-to-End learning for music audio*. Paper presented at IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), Florence, Italy, 2014.
- Downie, J. S. (2003). Music information retrieval. *Annual Review of Information Science and Technology*, 37, 295-340.
- Downie, J. S. (2004). The scientific evaluation of music information retrieval systems: foundations and the future. *Computer Music Journal*, 28(2), 12-23.
- Dubnov, S. & Assayag, G. (2012). *Music design with audio oracle using information rate*. In Musical Metacreation: Papers from the 2012 AIIDE Workshop. AAAI Technical Report WS-12-16 published by The AAAI Press, Palo Alto, California.
- Eghbal-zadeh, H., Lehner, B., Schedl, M., & Widmer, G. (2015). *I-vectors for timbre-based music similarity and music artist classification*. Paper presented at International Conference on Music Information Retrieval, Málaga, Spain.
- Eghbal-zadeh, H., Schedl, M., & Widmer, G. (2015). *Timbral modeling for music artist recognition using i-vectors*. Paper presented at 23rd European Signal Processing Conference (EUSIPCO), Nice, France, 1286–1290.
- Ellis, D.P.W. (1996). Prediction-driven computational auditory scene analysis. PhD thesis. Massachusetts Institute of Technology, Cambridge, MA, USA.
- Faraldo, Á., Jordà S., & Herrera P. (2017). *The house harmonic filler: Interactive exploration of chord sequences by means of an intuitive representation*. Paper



presented at 3rd International Conference on Technologies for Music Notation and Representation (TENOR-2017), A Coruña, Spain.

- Fernández-Tobías, I., Braunhofer, M., Elahi, M., Ricci, F. & Cantador, I. (2016) Alleviating the new user problem in collaborative filtering by exploiting personality information. *User Modeling and User-Adapted Interaction*, 26 (2-3), 221-255.
- Ferwerda, B., Graus, M.P., Vall, A., Tkalcic, M., Schedl, M. (2016). *The influence of users' personality traits on satisfaction and attractiveness of diversified recommendation lists*. Paper presented at 4th Workshop on Emotions and Personality in Personalized Systems (EMPIRE) 2016, Boston, MA, USA.
- Fiebrink, R. (2011). Real-time human interaction with supervised learning algorithms for music composition and performance. PhD thesis, Princeton University, Princeton, NJ, USA.
- Fiebrink, R., & Caramiaux, B. (2018). The machine learning algorithm as creative musical tool. In A. McLean and R. Dean (eds.), *The Oxford Handbook of Algorithmic Music*. Oxford University Press.
- Fields, B., Casey, M., Jacobson, K., & Sandler, M. (2008). *Do you sound like your friends? Exploring artist similarity via artist social network relationships and audio signal processing*. Paper presented at International Computer Music Conference (ICMC-2008), Belfast, UK.
- Fields, B. (2011). Contextualize your listening: the playlist as recommendation engine. PhD thesis, Goldsmiths University of London, London, UK.
- Fisher, A., & Sloutsky, V. M. (2005). When induction meets memory: evidence for gradual transition from similarity-based to category-based induction. *Child Development*, 76(3), 583-597.
- Fiske, H. (2008). *Understanding Musical Understanding: The Philosophy, Psychology, and Sociology of the Musical Experience*. Lewiston: Edwin Mellen Press.
- Foster, P., Mauch, M., & Dixon, S. (2014). Sequential complexity as a descriptor for musical similarity. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(12), 1965-1977.
- Fuhrmann, F., Herrera, P., Serra, X. (2009). Detecting solo phrases in music using spectral and pitch-related descriptors. *Journal of New Music Research*, 38(4), 343-356.
- Garner, W. R. (1974). *The processing of information and structure*. Potomac, MD: Lawrence Erlbaum.
- Gilda, S., Zafar, H., Soni, C. & Waghurdekar, K. (2017). *Smart music player integrating facial emotion recognition and music mood recommendation*. Paper presented at International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai, India, 154-158.
- Goldstone, R. L. (1994). The role of similarity in categorization: Providing a groundwork. *Cognition*, 52, 125-157.
- Goldstone, R. L. (1999). Similarity. In R.A. Wilson & F. C. Keil (Eds.), *The MIT encyclopedia of the cognitive sciences*. Boston, MA: MIT Press.

- Goldstone, R. L., Kersten, A., & Carvalho, P. F. (2017). Categorization and concepts. In J. Wixted (Ed.) *Stevens' Handbook of Experimental Psychology and Cognitive neuroscience*, Fourth Edition, Volume Three: Language & Thought (pp. 275-317) New Jersey: Wiley
- Gómez, E., & Herrera, P. (2004). *Automatic extraction of tonal metadata from polyphonic audio recordings*. Paper presented at 25th International AES Conference, London, UK.
- Gómez, E., & Herrera, P. (2006). *The song remains the same: identifying versions of the same piece using tonal descriptors*. Paper presented at 7th Intl. Conference on Music Information Retrieval, Victoria, Canada.
- Gómez-Marín, D., Jordà S., & Herrera P. (2017). *Drum rhythm spaces: from global models to style-specific maps*. Paper presented at 13th International Symposium on Computer Music Multidisciplinary Research (CMMR), Porto, Portugal.
- Goodman, N. (1972). Seven strictures on similarity. In N. Goodman (Ed.), *Project and problems*. Indianapolis, IN: The Bobbs-Merrill Company, Inc.
- Goto, M. (2001). *A predominant-F0 estimation method for CD recordings: MAP estimation using EM algorithm for adaptive tone models*. Paper presented at IEEE International Conference On Acoustics Speech And Signal Processing..
- Gouyon, F., & Herrera, P. (2003). *Determination of the meter of musical audio signals: Seeking recurrences in beat segment descriptors*. Paper presented at 114th Convention of the Audio Engineering Society, Amsterdam, The Netherlands.
- Grachten; M., Arcos, J. Ll.; López de Mántaras, R. (2006). A case based approach to expressivity-aware tempo transformation. *Machine Learning*, Volume 65, Number 2-3, p.411-437
- Griffin, G., Kim, Y. E., & Turnbull, D. (2010). *Beat-Sync-Mash-Coder: A web application for real-time creation of beat-synchronous music mashups*. Paper presented at ICASSP, Dallas, TX, USA, 437-440.
- Hamel, P. & Eck, D. (2010). *Learning features from music audio with deep belief networks*, Paper presented at International Conference of Music Information Retrieval, 339-344.
- Han, B., Rho, S., Jun, S. et al. (2010) Music emotion classification and context-based music recommendation. *Multimedia Tools and Applications*, 47(3), 433-460
- Han, Y., Kim, J., & Lee, K.. (2017). Deep convolutional neural networks for predominant instrument recognition in polyphonic music. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(1), 208-221.
- Hargreaves, D. J., & North, A. C. (1999). The functions of music in everyday life: Redefining the social in music psychology. *Psychology of Music*, 27(1), 71-83.
- Haro, M., Herrera, P. (2009). From low-level to song-level percussion descriptors of polyphonic music. Paper presented at 10th International Society for Music Information Retrieval Conference, Kobe, Japan.
- Hattwick, I., Beebe, P., Hale, Z., Wanderley, M. M., Leroux, P., & Marandola, F. (2013). *Unsounding objects: Audio feature extraction for control of sound synthesis in a*

*digital percussion instrument*. Paper presented at International Conference on New Interfaces for Musical Expression. London, England, 597-600.

- Hausfeld, L., Riecke, L., Valente, G., Formisano, E. (2018). Cortical tracking of multiple streams outside the focus of attention in naturalistic auditory scenes. *NeuroImage*, 181, 617-626.
- Hemery, E., & Aucouturier, J.-J. (2015). One hundred ways to process time, frequency, rate and scale in the central auditory system: a pattern-recognition meta-analysis. *Frontiers in Computational Neuroscience*, 9, 80.
- Henry, M.H., Taylor, S. & Garrison, C. (2018). *DeepJ: Style-specific music generation*, Paper presented at IEEE ICSC Conference 2018, 377-382.
- Herremans, D., Chuan, C. H. & Chew, E. (2017). A functional taxonomy of music generation systems. *ACM Comput. Surv.* 50(5), 69:1-69:30.
- Herrera, P., Amatriain, X., Batlle, E., & Serra, X. (2000). *Towards instrument segmentation for music content description: A critical review of instrument classification techniques*. Paper presented at International Symposium on Music Information Retrieval, Plymouth, MA, USA.
- Herrera, P., Bello, J. P., Widmer, G., Sandler, M., Celma, O., Vignoli, F. et al. (2005). *SIMAC: Semantic interaction with music audio contents*. Paper presented at 2nd European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies, London, UK.
- Herrera, P., Dehamel, A., & Gouyon, F. (2003). *Automatic labeling of unpitched percussion sounds*. Paper presented at 114th Convention of the Audio Engineering Society, Amsterdam, The Netherlands.
- Herrera, P., Peeters, G., & Dubnov, S. (2003). Automatic classification of musical instrument sounds. *Journal of New Music Research*, 32(1), 3-21.
- Herrera, P., Resa, Z., & Sordo, M. (2010). *Rocking around the clock eight days a week: an exploration of temporal patterns of music listening*. Paper presented at 1st Workshop On Music Recommendation And Discovery (WOMRAD), Barcelona, Spain.
- Herrera, P., Sandvold, V., Gouyon, F. (2004). *Percussion-related semantic descriptors of music*. Paper presented at 25th Interantional Conference of the Audio Engineering Society, London, United Kingdom.
- Herrera, P., Serra, X., & Peeters, G. (1999a). *A proposal for the description of audio in the context of MPEG-7*. Paper presented at 1st International Workshop on Content-Based Multimedia Indexing (CBMI), Toulouse, France.
- Herrera, P., Serra, X., & Peeters, G. (1999b). *Audio descriptors and descriptor schemes in the context of MPEG-7*. Paper presented at International Computer Music Conference, ICMA, Beijing, China, 581-584.
- Herrera, P., Serrà, J., Laurier, C., Gaus, E., Gómez, E., Serra, X. (2009). *The discipline formerly known as MIR*. Paper presented at International Society for Music Information Retrieval (ISMIR) Conference, special session on The Future of MIR (fMIR), Kobe, Japan.

- Herrera, P., Yeterian, A., & Gouyon, F. (2002). Automatic classification of drum sounds: a comparison of feature selection methods and classification techniques. Anagnostopoulou, C., Ferrand, M., and Smaill, A. *Music and Artificial Intelligence: Second International Conference Proceedings*. [2445], 69-80. Berlin, Springer. Lecture Notes in Computer Science.
- Herrera-Boyer, P., Klapuri, A., & Davy, M. (2006). Automatic classification of pitched musical instrument sounds. In A. Klapuri & M. Davy (Eds.), *Signal processing methods for music transcription* (pp. 163-200). New York: Springer.
- Hiller, L., & Isaacson, L. (1959). *Experimental music: Composition with an electronic computer*. New York: McGraw-Hill Book Company, Inc.
- Hirai, T., Doi, H., & Morishima, S. (2015). *MusicMixer: Computer-aided DJ system based on an automatic song mixing*. Paper presented at 12th International Conference on Advances in Computer Entertainment Technology. ACM, 41.
- Hu, X., Downie, J. S., & Ehmann, A. F. (2006). *Exploiting recommended usage metadata: Exploratory analyses*. Paper presented at International Conference on Music Information Retrieval, Victoria, Canada.
- Hu, Y. & Ogihara, M. (2011). *NextOne Player: A music recommendation system based on user behavior*. Paper presented at 12th International Society for Music Information Retrieval Conference. Miami (Florida), USA. 103-108.
- Humphrey, E. J., Bello, J. P., & LeCun, Y. (2012). *Moving beyond feature design: Deep architectures and automatic feature learning in music informatics*. Paper presented at International Society for Music Information Retrieval Conference (ISMIR), Porto, Portugal, 403–408.
- Humphrey, E.J., Bello, J.P., LeCun, Y. (2013). Feature learning and deep architectures: New directions for music informatics, *Journal of Intelligent Information Systems* 41 (3), 461-481.
- Humphrey, E.J., Turnbull, D., & Collins, T. (2013). *A brief review of creative MIR*. Paper presented at ISMIR 2013 Late-Breaking News and Demos, Curitiba, Brazil.
- Hyung, Z., Lee, K. & Lee, K. (2014). Music recommendation using text analysis on song requests to radio stations. *Expert Systems with Applications*, 41(5), 2608-2618.
- Imbir, K., & Gołab, M. (2016). Affective reactions to music: Norms for 120 excerpts of modern and classical music. *Psychology of Music*, 45 (3), 432-449
- Ishizaki, H., Hoashi, K., & Takishima, Y. (2009). *Full-automatic DJ mixing system with optimal tempo adjustment based on measurement function of user discomfort*. Paper presented at ISMIR, Kobe, Japan, 135–140.
- Jehan, T. (2005). *Creating music by listening*. PhD thesis. Massachusetts Institute of Technology, Cambridge, MA, USA.
- Jones, M. C., Downie, J. S., & Ehmann, A. F. (2007) *Human similarity judgments: implications for the design of formal evaluations*. Paper presented at 8th International Conference on Music Information Retrieval (ISMIR), Vienna, Austria, 539-542.

- Kallinen, K., & Ravaja, N. (2004). The role of personality in emotional responses to music: Verbal, electrocortical and cardiovascular measures. *Journal of New Music Research*, 33(4), 399-409.
- Kamehkhosh, I., Jannach, D., and Lerche, L. (2016) *Personalized next-track music recommendation with multi-dimensional long-term preference signals*. Paper presented at IFUP Workshop at UMAP '16, Halifax, Canada.
- Kaminskas, M., Ricci, F. (2009). Contextual music information retrieval and recommendation: state of the art and challenges. *Computer Science Review*. 6, 89–119
- Kaminskas, M., Ricci, F., & Schedl, M. (2013). *Location-aware music recommendation using auto-tagging and hybrid matching*. Paper presented at 7th ACM Conference on Recommender Systems (RecSys). Hong Kong, China.
- Karmaker, D., Imran A., Mohammad, N., Islam, M., Mahbub, N. (2018). An automated music selector derived from weather condition and its impact on human psychology. *GSTF Journal on Computing*, 4(3).
- Kereliuk, C., Sturm, B.L., & Larsen, J. (2015). Deep learning and music adversaries. *IEEE Transactions on Multimedia*, 17, 2059-2071.
- Kim, Y.E. (2003). Singing voice analysis/synthesis. PhD thesis. Massachusetts Institute of Technology, Cambridge, MA, USA.
- Klapuri, A., & Davy, M. (2006). Signal processing methods for music transcription. New York: Springer.
- Knees, P. & Schedl, M. (2013). A survey of music similarity and recommendation from music context data. *ACM Trans. Multimedia Comput. Commun. Appl.* 10, 1, 1-21.
- Knees, P. & Schedl, M. (2016). Music similarity and retrieval: An introduction to audio- and web-based strategies. Springer.
- Knees, P., Schedl, M., & Pohle, T. (2008). *A deeper look into webbased classification of music artists*. Paper presented at 2nd Workshop on Learning the Semantics of Audio Signals, Paris, France.
- Knox, D., Beveridge, S., Mitchell, L., & MacDonald, R. (2011). Acoustic analysis and mood classification of pain-relieving music. *Journal of the Acoustical Society of America*, 130(3), 1673-1682.
- Koelsch, S., Kasper, E., Sammler, D., Schulze, K., Gunter, T., & Friederici, A. (2004). Music, language and meaning: Brain signatures of semantic processing. *Nature Neuroscience*, 7(3), 302-307.
- Koelsch, S., & Siebel, W. A. (2005). Towards a neural basis of music perception. *Trends in Cognitive Sciences*, 9(12), 578-584.
- Krause, A. E., & North, A. C. (2018). 'Tis the season: Music-playlist preferences for the seasons. *Psychology of Aesthetics, Creativity, and the Arts*, 12(1), 89-95.
- Kubovy, M. (1981). Integral and separable dimensions and the theory of indispensable attributes. In M. Kubovy & J. Pomerantz (Eds.), *Perceptual organization*. Hillsdale, NJ: Erlbaum.

- Kuhn, Thomas S. (2012). The structure of scientific revolutions. 50th anniversary. Ian Hacking (intro.) (4th ed.). University of Chicago Press.
- Lamere, P. (2008). Social tagging and music information retrieval. *Journal of New Music Research*, 37(2), 101-114.
- Lamont, A., & Dibben, N. (2001). Motivic structure and the perception of similarity. *Music Perception*, 18, 245-274.
- Laurier, C., Grivolla, J., & Herrera, P. (2008). *Multimodal music mood classification using audio and lyrics*. Paper presented at International Conference on Machine Learning and Applications, San Diego, USA.
- Laurier, C., Herrera, P. (2008). *Mood Cloud : A real-time music mood visualization tool*. Paper presented at Computer Music Modeling and Retrieval Conference, Copenhagen, Denmark.
- Laurier, C., Meyers, O., Serrà, J., Blech, M., Herrera, P., & Serra, X. (2010). Indexing music by mood: Design and integration of an automatic content-based annotator. *Multimedia Tools and Applications*, 48(1), 161-184.
- LeCun, Y.; Bottou, L., Bengio, Y. & Haffner, P. (1998). Gradient-based learning applied to document recognition, *Proceedings of the IEEE*, 86, 2278-2324.
- Lee, H. Pham, P., Largman, Y. & Ng, A. Y. (2009). *Unsupervised feature learning for audio classification using convolutional deep belief networks*. Paper presented at Advances in neural information processing Systems conference, Vancouver, B.C., Canada, 1096-1104.
- Lee, J., & Lee, J. (2006). *Music for my mood: a music recommendation system based on context reasoning*. Paper presented at Smart Sensing and Context, Enschede, The Netherlands, 190-203.
- Lee, J. H., & Downie, J. S. (2004). *Survey of music information needs, uses, and seeking behaviours: preliminary findings*. Paper presented at 5th International Conference on Music Information Retrieval, Barcelona, Spain.
- Lee J., & Nam, J. (2017). Multi-level and multi-scale feature aggregation using pre-trained convolutional neural networks for music auto-tagging. *IEEE Signal Processing Letters*, 24 (8), 1208-1212.
- Lee, K., Choi, K., and Nam, J. (2018). Revisiting singing voice detection: a quantitative review and the future outlook. Paper presented at 19th International Society for Music Information Retrieval Conference (ISMIR), Paris, France.
- Lerch, A. (2012). An introduction to audio content analysis. New York, Wiley.
- Lesaffre, M. (2006). Music information retrieval: Conceptual framework, annotation and user behaviour, PhD Thesis, Ghent University, Ghent, Belgium.
- Lesaffre, M., Leman, M., Tanghe, K., De Baets, B., De Meyer, H., & Martens, J. P. (2003). *User-dependent taxonomy of musical features as a conceptual framework for musical audio-mining technology*. Paper presented at Stockholm Music Acoustics Conference (SMAC 03).

- Levy, M. A & Sandler, M. (2007). *A semantic space for music derived from social tags*. Paper presented at 8th International Conference on Music Information Retrieval (ISMIR '07), pp. 411–416, Vienna, Austria, September 2007.
- Levy, M., & Sandler, M. (2008). Learning latent semantic models for music from social tags. *Journal of New Music Research*, 37, 137-150.
- Libeks, J., & Turnbull, D. (2010). *Exploring artist image using content-based analysis of promotional photos*. Paper presented at International Computer Music Conference, Utrecht, Netherlands.
- Libeks, J., & Turnbull, D. (2011). You Can Judge an Artist by an Album Cover: Using Images for Music Annotation. *IEEE MultiMedia*, 18, 30-37.
- Liu, K., & Reimer, R. A. (2008). *Social playlist: enabling touch points and enriching ongoing relationships through collaborative mobile music listening*. Paper presented at MobileHCI '08, Barcelona, Spain, 403-406.
- Logan, B., Kositsky, A., & Moreno, P. J. (2004). *Semantic analysis of song lyrics*. Paper presented at IEEE International Conference on Multimedia and Expo (ICME), 2, 827-830.
- Man, B. D. & Reiss, J. D. (2013). A semantic approach to autonomous mixing, *Journal of the Art of Record Production*, 8.
- Man, B. D., Reiss, J. D., & Stables, R. (2017). *Ten years of automatic mixing*. Paper presented at 3rd Workshop on Intelligent Music Production, Salford, UK.
- Martin, K. D. (1999). Sound-source recognition: A theory and computational model. PhD thesis. Massachusetts Institute of Technology, Cambridge, MA, USA.
- Martin, K. D., Scheirer, E. D., & Vercoe, B. L. (1998). *Musical content analysis through models of audition*. Paper presented at ACM Multimedia Workshop on Content-Based Processing of Music, Bristol, UK.
- Masahiro, N., Takaesu, H., Demachi, H., Oono, M., & Saito, H. (2008). Development of an automatic music selection system based on runner's step frequency. Paper presented at 9th International Conference on Music Information Retrieval, Philadelphia, PA, USA, 193-198.
- Mayer, R., Rauber, A. (2010) Multimodal aspects of music retrieval: audio, song lyrics - and beyond?, in Z. B. Raś, & A. A. Wierzchowska (eds.) *Advances in Music Information Retrieval*, Springer-Verlag, Berlin, 333-363.
- McFee, B (2012). More like this: machine learning approaches to music similarity. PhD Thesis, Univ. of California, San Diego, USA.
- McFee, B., Humphrey, E. J., & Bello, J. P. (2015). *A software framework for musical data augmentation*. Paper presented at International Conference on Music Information Retrieval, Málaga, Spain, 248–254.
- McVicar, M., Fukayama, S., Goto, M. (2014) *AutoLeadGuitar: Automatic generation of guitar solo phrases in the tablature space*. Paper presented at IEEE 12th International Conference on Signal Processing (ICSP 2014) HangZhou, China, 599–604.

- Mechtley, B.M. (2013). Techniques for soundscape retrieval and synthesis. PhD Thesis, Arizona State University, USA.
- Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, 100(2), 254-278.
- Mesaros, A., Heittola, T., and Palomäki, K. (2013). *Analysis of acoustic-semantic relationship for diversely annotated real-world audio data*. Paper presented at, IEEE Int. Conf. on Acoustics, Speech and Signal Processing ICASSP 2013, San Francisco, CA, USA .
- Miller, G. A. (1995). WordNet: a lexical database for english. *Communications of the ACM*, 38(11), 39-41.
- Müller, M. (2015). Fundamentals of music processing. Berlin, Springer.
- Müller, M., & Driedger, J.. (2012). Data-driven sound track generation. In M.Müller (ed.), *Multimodal Music Processing*. Springer, 175–194.
- Müller, M., Konz, V., Scharfstein, A., Ewert, S., & Clausen, M. (2009). *Towards automated extraction of tempo parameters from expressive music recordings*. Paper presented at 10th International Conference on Music Information Retrieval, 69-72)
- Müller, M., Weiss, C., Balke, S. (2017). *Deep Neural Networks in MIR*. Tutorial Automatisierte Methoden der Musikverarbeitung 47. Jahrestagung der Gesellschaft für Informatik.
- Murthy Y. V. S. & Koolagudi, S.G. (2018). Content-based music information retrieval (CB-MIR) and its applications toward the music industry: a review. *ACM Comput. Surv.*, 51 (3), 1-46.
- Nack, F. (2004). The future in digital media computing is meta. *IEEE Multimedia*, 11(2), 10-13.
- Nagel, F. (2007). Psychoacoustical and psychophysiological correlates of the emotional impact and the perception of music. PhD Thesis, Hannover University of Music and Drama, Hannover, Germany.
- Navarro, F. L., Herrera P., & Gómez, E. (2014). *There are places I remember: Personalized automatic creation of music playlists for Alzheimer's patients*. Paper presented at The Neurosciences and Music V, Dijon, France.
- Oliver, N., & Kreger-Stickles, L. (2006). *PAPA: Physiology and purpose-aware automatic playlist generation*. Paper presented at 7th International Conference on Music Information Retrieval, Victoria, Canada.
- Oramas, S., Nieto, O., Barbieri, F., & Serra, X. (2017). *Multi-label music genre classification from audio, text and images using deep features*. Paper presented at 18th International Conference on Music Information Retrieval, Suzhou, China.
- Pachet, F., Roy, P. and Ghedini, F. (2013). *Creativity through style manipulation: the Flow Machines project*. Paper presented at Marconi Institute for Creativity Conference (MIC 2013) vol. 80, Bologna, Italy.
- Pachet, F., Roy, P., Moreira, J. & d'Inverno, M. (2013). *Reflexive loopers for solo musical improvisation*. Paper presented at SIGCHI ACM Conference on Human Factors in Computing Systems, Paris, France, 2205-2208.



- Pachet, F., Westerman, G., & Laigre, D. (2001). *Musical data mining for electronic music distribution*. Paper presented at 1st Conference on Web Delivering of Music (Wedelmusic), Florence, Italy.
- Pampalk, E. (2006). Computational models of music similarity and their application in music information retrieval. PhD thesis, OFAI, Vienna, Austria.
- Pampalk, E., Flexer, A., & Widmer, G. (2005). *Improvements of audio-based music similarity and genre classificaton*. Paper presented at 6th International Conference on Music Information Retrieval, London, UK.
- Pampalk, E., Herrera, P., Goto, M. (2008). Computational models of similarity for drum samples. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2), 408-423.
- Pampalk, E., Hlavac, P., Herrera, P. (2004). *Hierarchical organization and visualization of drum sample libraries*. Paper presented at the 7th International Conference on Digital Audio Effects, Naples, Italy.
- Pampalk, E., Pohle, T. and Widmer, G. (2005). *Dynamic playlist generation based on skipping behaviour*. Paper presented at 6th International Conference on Music Information Retrieval, London, UK, 634-637.
- Pardo, B., Little, D., and Gergle, D. (2012). *Building a personalized audio equalizer interface with transfer learning and active learning*. Paper presented at 2nd International ACM Workshop on Music Information Retrieval with User-centered and Multimodal Strategies, Nara, Japan.
- Parera, J. (2016). DJ Codo Nudo: a novel method for seamless transition between songs for electronic music. MSc thesis, Universitat Pompeu Fabra, Barcelona, Spain.
- Park, H.-S., Yoo, J.-O., & Choo, S. B. (2006). *A Context-aware music recommendation system using fuzzy bayesian networks with utility theory*. Paper presented at 3rd international conference on Fuzzy Systems and Knowledge Discovery (FSKD'06), 970-979
- Park, J., Kim, D., Lee, J., Kum, S, Nam, J. (2018). *A hybrid of deep audio feature and i-vector for artist recognition*. Paper presented at Joint Workshop on Machine Learning for Music, the 34th International Conference on Machine Learning (ICML), Sydney, Australia.
- Peeters, G. (2004). A large set of audio features for sound description (similarity and classification) in the CUIDADO project [Technical report].
- Peeters, G. (2006). *Musical key estimation of audio signal based on hidden Markov modeling of chroma vectors*. Paper presented at International Conference on Digital Audio Effects (DAFx'06), Montreal, Quebec, Canada,127-131.
- Peeters G., Cornu, F., Doukhan, D., Marchetto, E., Mignot, R., Perros, K., & Regnier, L. (2015). *When audio features reach machine learning*. Paper presented at Machine Learning for Music Discovery Workshop at the 32nd International Conference on Machine Learning, Lille, France, 2015.
- Peeters, G. & Deruty, E. (2010). Sound indexing using morphological description. *IEEE Transactions on Audio, Speech and Language Processing*, 18 (3), 675-687.

- Peeters, G., McAdams, S., & Herrera, P. (2000). *Instrument sound description in the context of MPEG-7*. Paper presented at International Computer Music Conference, Berlin, Germany, 166-169.
- Peeters, G., Fort, K. (2012). *Towards a (better) definition of annotated MIR corpora*. Paper presented at International Society for Music Information Retrieval Conference (ISMIR), Porto, Portugal.
- Pettijohn II, T.F., Williams, G.M. & Carter, T.C. (2010). Music for the seasons: Seasonal music preferences in college students. *Current Psychology*, 29(4), 328-345.
- Plumbley M. D., Samer A. Abdallah, S., Bello J. P., Davies M. E., Monti G., and Sandler M. B., (2002) Automatic music transcription and audio source separation. *Cybernetics and Systems*, 33(6): 603-627.
- Pohle, T., Schnitzer, D., Schedl, M., Knees, P., & Widmer, G. (2009). *On rhythm and general music similarity*. Paper presented at 10th International Society for Music Information Retrieval Conference, Utrecht, The Netherlands, 525-530.
- Pons, J., Slizovskaia, O., Gómez-Gutiérrez, E., Serra, X. (2017). *Timbre analysis of music audio signals with convolutional neural networks*. Paper presented at European Signal Processing Conference (EUSIPCO), Kos, Greece, 2813-2819.
- Puvvada, K. C., & Simon, J. Z. (2017). Cortical representations of speech in a multitalker auditory scene. *J. Neuroscience*, 37, 9189-9196,
- Rentfrow, P. J., Goldberg, L. R., & Levitin, D. J. (2011). The structure of musical preferences: a five-factor model. *Journal of Personality and Social Psychology*, 100(6), 1139–1157.
- Rentfrow, P. J., & Gosling, S. D. (2003). The do re mi's of everyday life: The structure and personality correlates of music preferences. *Journal of Personality and Social Psychology*, 84, 1236-1256.
- Ricard, J., & Herrera, P. (2003). *Using morphological description for generic sound retrieval*. Paper presented at 4th International Conference on Music Information Retrieval, Baltimore, Maryland, USA.
- Ricard, J., & Herrera, P. (2004). *Morphological sound description: Computational model and usability evaluation*. Paper presented at 116th Convention of the Audio Engineering Society, Berlin.
- Richard, G., Sundaram, S., & Narayanan, S. (2013). An overview on perceptually motivated audio indexing and classification. *Proceedings of the IEEE*, vol. 101(9), 1939-1954.
- Rocamora, M., Herrera, P. (2007). *Comparing audio descriptors for singing voice detection in music audio files*. Paper presented at 11th Brazilian Symposium on Computer Music, San Pablo, Brazil.
- Roma, G., Herrera, P., Serra, X. (2009). *Freesound Radio: supporting music creation by exploration of a sound database*. Paper presented at Computational Creativity Support Workshop CHI09, Boston, MA, USA.
- Roma G., & Serra, X. (2015). *Music Performance by Discovering Community Loops*. Paper presented at 1st Web Audio Conference, Paris, France.

- Román Echeverri, C. G., & Herrera P. (2013). Automatic description of music for analyzing music productions: a case study in detecting mellotron sounds in recordings. *Journal on the Art of Record Production*, 8. ISSN: 1754-9892
- Roy, P., Perez, G., Régini, J.C., Papadopoulos, A., Pachet, F. & Marchini, M. *enforcing structure on temporal sequences: the allen constraint*. Paper presented at 22nd International Conference on Principles and Practice of Constraint Programming – CP 2016.
- Sakellariou, J., Tria, F., Loreto, V. and Pachet, F. (2017). Maximum entropy models capture melodic styles. *Scientific Reports*, 7(9172)
- Salselas, I., & Herrera P. (2011). Music and speech in early development: automatic analysis and classification of prosodic features from two Portuguese variants. *Journal of Portuguese Linguistics*, 9/10, 11-36.
- Sandvold, V., Gouyon, F. & Herrera, P. (2004). *Drum sound classification in polyphonic audio recordings using localized sound models*. Paper presented at 5th International Conference on Music Information Retrieval, Barcelona, Spain.
- Sandvold, V., & Herrera, P. (2005). *Towards a semantic descriptor of subjective intensity in music*. Paper presented at International Computer Music Conference, Barcelona, Spain.
- Sarasúa, Á., Laurier C., & Herrera P. (2012). *Support Vector Machine active learning for music mood tagging*. Paper presented at 9th International Symposium on Computer Music Modeling and Retrieval (CMMR). 518-525.
- Schaeffer, P. (1966). *Traité des objets musicaux*. Paris: Éditions Du Seuil. (English translation: Schaeffer, P. (2017). *Treatise on Musical Objects: An Essay across Disciplines*. University of California Press)
- Schankler, I., Chew, E., & François, A. (2014). Improvising with digital auto-scaffolding: how Mimi changes and enhances the creative process. In Lee N. (ed.) *Digital Da Vinci*. Springer, New York, 99–125.
- Schedl, M. (2008). *Automatically extracting, analyzing, and visualizing information on music artists from the world wide web*. PhD thesis, Johannes Kepler Universität, Linz, Austria.
- Schedl, M. (2013). *Ameliorating music recommendation: Integrating music content, music context, and user context for improved music retrieval and recommendation*. Paper presented at 11th International Conference on Advances in Mobile Computing & Multimedia (MoMM 2013), Vienna, Austria, December 2013.
- Schedl, M. (2017). Investigating country-specific music preferences and music recommendation algorithms with the LFM-1b dataset. *International Journal of Multimedia Information Retrieval*, 6(1), 71–84.
- Schedl, M., Breitschopf, G., & Ionescu, B. (2014). *Mobile Music Genius: Reggae at the beach, metal on a friday night?* Paper presented at 4th ACM International Conference on Multimedia Retrieval (ICMR). Glasgow, UK.
- Schedl, M., Flexer, A. & Urbano, J. (2013). The neglected user in music information retrieval research, *Journal of Intelligent Information Systems*, 41(3), 523–539.

- Schedl, M., Gómez, E. & Urbano, J. (2014). Music information retrieval: Recent developments and applications. *Found. Trends Inform. Retrieval.*, 8(2–3), 127–261.
- Schedl, M., & Knees, P. (2009). *Context-based music similarity estimation*. Paper presented at 3rd International Workshop on Learning the Semantics of Audio Signals (LSAS 2009), Graz, Austria.
- Schedl, M., & Knees, P. (2011). *Personalization in multimodal music retrieval*. Paper presented at 9th international conference on Adaptive Multimedia Retrieval: large-scale multimedia retrieval and evaluation (AMR'11), 58-71.
- Schedl, M., Stober, S., Liem, C., Gouyon, F., Orio, N., & Gómez, E. (2012). User-aware music retrieval and recommendation. In M. Müller, M. Goto, & M. Schedl (Eds.), *Multimodal Music Processing* (pp. 1-16). Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Germany: Dagstuhl Publishing.
- Schedl, M., Knees, P., Pohle, T., & Widmer, G. (2008). *Towards an automatically generated music information system via web content mining*. Paper presented at 30th European Conference on Information Retrieval (ECIR'08), Glasgow, Scotland.
- Scheirer, E. (2000). Music listening systems. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, USA.
- Schindler, A. & Rauber, A. (2015). *An audio-visual approach to music genre classification through affective color features*. Paper presented at 37th European Conference on Information Retrieval (ECIR'15), Vienna, Austria.
- Schlüter, J. & Grill, T. (2015). *Exploring data augmentation for improved singing voice detection with neural networks*. Paper presented at 16th International Society for Music Information Retrieval Conference, 121–126.
- Schmidt, E. M. & Kim, Y. E. (2013). *Learning rhythm and melody features with deep belief networks*. Paper presented at International Society for Music Information Retrieval Conference, Curitiba, Brazil, 2013.
- Schubert, E. (2002). Continuous measurement of self-report emotional response to music. In P.N. Juslin & J. A. Sloboda (Eds.), *Music and Emotion: Theory and Research* (pp. 393-414). Oxford, UK: Oxford University Press.
- Schwarz, D., Beller, G., Verbrugge, B., & Britton, S.. (2006). *Real-time corpus-based concatenative synthesis with CataRT*. Paper presented at 9th International Conference on Digital Audio Effects, 279-282.
- Serra, X. A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition (1989). PhD. Thesis, Stanford University, Stanford, CA, USA.
- Serra, X. Bonada, J. Herrera, P. Loureiro, R. (1997). *Integrating complementary spectral models in the design of a musical synthesizer*. Paper presented at International Computer Music Conference 1997, Thessaloniki, Greece.
- Serra, X., Magas, M., Benetos, E., Chudy, M., Dixon, S., Flexer, A., Gómez, E., Gouyon, F., Herrera, P., Jorda, S., Paytuvi, O., Peeters, G., Schlüter, J., Vinet, H., Widmer, G. (2013). Roadmap for Music Information ReSearch, G. Peeters (ed.), Creative Commons BY-NC-ND 3.0 license (ISBN: 978-2-9540351-1-6)

- Serrà, J., Zanin, M., Herrera, P., & Serra, X. (2012). Characterization and exploitation of community structure in cover song networks. *Pattern Recognition Letters*, 33(1), 1032-1041.
- Shavitt, Y., Weinsberg, U. (2009). *Songs clustering using peer-to-peer co-occurrences*. Paper presented at IEEE ISM: AdMIRe, San Diego, CA.
- Smaragdís. P. (2001). Redundancy reduction for computational audition, a unifying approach. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, USA.
- Smith, L. B. (1989). From global similarity to kinds of similarity: The construction of dimensions in development. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 146-178). Cambridge: Cambridge University Press.
- Snyder, B. (2000). *Music and memory*. Cambridge, MA: MIT Press.
- Sordo, M., Laurier C., & Celma Ò. (2007). *Annotating Music Collections: How content-based similarity helps to propagate labels*, Paper presented at 8th International Conference on Music Information Retrieval, Vienna, Austria.
- Stober, S. (2011). Adaptive methods for user-centered organization of music collections, PhD Thesis, Otto-von-Guericke-Universität, Magdeburg, Germany.
- Streich, S. & Herrera, P. (2005). *Detrended fluctuation analysis of music signals: danceability estimation and further semantic characterization*. Paper presented at 118th Audio Engineering Society Convention, Amsterdam, The Netherlands.
- Streich, S. & Herrera, P. (2006). *Algorithmic prediction of music complexity judgements*. Paper presented at 9th International Conference on Music Perception and Cognition, Bologna, Italy.
- Sturm, B. L. (2012). *An analysis of the GTZAN music genre dataset*. Paper presented at the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies (MIRUM '12). ACM, New York, NY, USA, 7-12.
- Sturm, B. L.. (2014). The state of the art ten years after a state of the art: Future research in music information retrieval. *Journal of New Music Research*, 43(2), 147-172.
- Sturm, B.L.. (2015). A simple method to determine if a music information retrieval system is a “horse”. *IEEE Trans. Multimedia* 16 (6), 1636-1644
- Sturm, B.L. & Collins, N. (2014). *Four challenges for music information retrieval researchers*. Paper presented at DMRN+9: Digital Music Research Network One-day Workshop 2014.
- Sturm, B.L., Bardeli, R., Langlois, T., Emiya, V. (2014). *Formalizing the problem of music description*. Paper presented at 15th International Conference on Music Information Retrieval, Taipei, Taiwan.
- Thaut, M. H., & Davis, W. B. (1993). The influence of subject-selected versus experimenter-chosen music on affect, anxiety, and relaxation. *Journal of Music Therapy*, 30(210), 223.
- Tillmann, B., & Bharucha, J. J. (2000). Implicit learning of tonality: A self-organizing approach. *Psychological Review*, 107(4), 885-913.

- Tintarev, N., & Masthoff, J. (2007). *Effective explanations of recommendations: user-centered design*. Paper presented at ACM conference on recommender Systems, 153–156.
- Tversky, A. (1977). Features of similarity. *Psychological Review*, 84, 327-352.
- Uhlich, S. Porcu, M., Giron, F., Enenkl, M., Kemp, T., Takahashi, N. , Mitsufuji, Y. (2017). *Improving music source separation based on deep neural networks through data augmentation and network blending*. Paper presented at IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, 261-265.
- Uitdenbogerd, A., & Van Steelant, D. (2002). *A review of factors affecting music recommender success*. Paper presented at 3rd International Conference on Music Information Retrieval, Paris, France, 204-208.
- Unknown author (2011). News on IEEE ICME and AdMIRe. *Computer Music Journal*, 35(4), 10.
- Urbano, J. (2013). Evaluation in audio music similarity, Ph.D. thesis, Universidad Carlos III de Madrid, Madrid, Spain.
- Urbano, J. & Schedl, M. (2013). Minimal test collections for low-cost evaluation of Audio Music Similarity and Retrieval systems, *International Journal of Multimedia Information Retrieval*, 2(1), 59-70.
- Van Balen, J. M. H. (2016) Audio description and corpus analysis of popular music. PhD thesis, Utrecht University, Utrecht, The Netherlands.
- Vall, A., Quadrana, M., Schedl, M., Widmer, G., & Cremonesi, P. (2017). *The importance of song context in music playlists*. Paper presented at the Poster Track of the 11th ACM Conference on Recommender Systems (RecSys), RecSys '17, Como, Italy, 2017.
- Vinet, H., Herrera, P., Pachet, F. (2002). *The CUIDADO project*. Paper presented at 3rd International Conference on Music Information Retrieval, Paris, France.
- Wack, N., Laurier, C., Meyers, O., Marxer, R., Bogdanov, D., Serrà, J. et al. (2010). *Music classification using high-level models*. Summary of MIREX 2010 submission, 11th International Conference on Music Information Retrieval, Utrecht, The Netherlands.
- Wang, X., Rosenblum, D., Wang, Y. (2012). *Context-aware mobile music recommendation for daily activities*. Paper presented at 20th ACM International Conference on Multimedia. ACM, Nara, Japan, 99–108.
- Ward, D., Mason, R. D., Kim, C., Stöter, F.-R., Liutkus, A. and Plumbley, M. D. (2018). *SiSEC 2018: State of the art in musical audio source separation - subjective selection of the best algorithm*. Paper presented at 4th Workshop on Intelligent Music Production (WIMP) 2018.
- Weih, C., Jannach, D., Vatulkin, I., Rudolph, G. (2016). Music data analysis: Foundations and applications. Chapman & Hall/CRC Computer Science & Data Analysis.

- Whitman, B. (2005). Learning the meaning of music. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, USA.
- Whitman, B., & Lawrence, S. (2002). *Inferring descriptions and similarity for music from community metadata*. Paper presented at International Computer Music Conference, 591-598.
- Wiggins, G. A. (2006). A preliminary framework for description, analysis and comparison of creative systems, *Knowledge-Based Systems*, 19(7), 449–458.
- Wiggins, G. A. (2009). *Semantic gap?? Schemantic schmap!! Methodological considerations in the scientific study of music*. Paper presented at Workshop on Advances in Music Information Research (AdMIRe-2009).
- Witten, I., Frank, E., Hall, M., and Pal, C. J. (2016). Data mining: Practical machine learning tools and techniques (4th ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
- Woods, K.J.P., & McDermott, J. H. (2015) Attentive tracking of sound sources. *Current Biology*, 25(17), 2238-2246.
- Xambó, A., Roma, G., Lerch, A., Barthelet, M., & Fazekas, G. (2018). *Live repurposing of sounds: MIR explorations with personal and crowdsourced databases*. Paper presented at NIME 2018: New Interfaces for Musical Expression, 364-369. Blacksburg, Virginia, USA.
- Zils, A., & Pachet, F. (2001). *Musical mosaicing*. Paper presented at Conference on Digital Audio Effects (DAFX-01), Limerick, Ireland.
- Zils, A., & Pachet, F. (2004). *Automatic extraction of music descriptors from acoustic signals using EDS*. Paper presented at 116th AES Convention, Berlin, Germany.