Issues on Retrieval of Sound Effects in Large Collaborative Databases

Elena Martínez

Master Thesis submitted in partial fulfillment
of the requirements for the degree:
Master in Information, Communication and Audiovisual Media Technologies

Supervisor:
Dr. Xavier Serra

Department of Information and Communication Technologies
Universitat Pompeu Fabra
Barcelona, Spain
September 2008

Elena Martínez

Abstract

New ways of producing information, knowledge, and culture through social, rather than proprietary relations, are probably reponsible for the recent proliferation of online communities. Many of these communities aim to collaboratively create large multimedia databases. The context in which these sites are growing and many of their important aspects are presented and discussed in this work. From all of them, the retrieval issues of sound effects have been selected as main focus of the thesis. Specifically, aspects concerning the annotation of such large databases by means of collaborative tagging, and others dealing with the study of alternative ways to retrieve audio content, such as sound search by phonetic similarity.

The collaborative sound database *Freesound.org* has been chosen for the experiments. First of all, an study about issues such as how users annotate the sounds in the database, have been conducted, detecting some well–known problems in collaborative tagging, such as polysemy, synonymy, and the scarcity of the existing annotations. Then, a subset of sounds rarely or scarcely tagged were selected for the experiments, where a content–based audio similarity system aiming at automatically enhancing these annotations have been proposed. The reported results show that 77% of the test collection were enhanced using the recommended tags, with a high agreement among the subjects.

We have also studied alternative interfaces that allow different criteria when browsing and retrieving these databases, with the motivation that such large repositories may have different levels of descriptions, and thus different ways of browsing may be allowed. The proposed system attempts on extracting automatically timbre information at the phoneme level, by means of phonetic similarity. Hence, the keyword-based search can be complemented, allowing at the same time a more natural retrieval, since given that onomatopoeic representation of sounds are very common ways to describe them. Promising results are reported, and even still further improvements have to be considered in order to extend it to more sound categories with successful performance.

Acknowledgments

I would like to express my gratitude to the constant support received from my supervisor, Dr. Xavier Serra, who gave me the opportunity to join the Music Technology Group, as well as to take part in the research carried out in the Freesound Project. Sincerely, it has been a pleasure to take part of such brilliant group of nice people.

I would also like to thank Oscar Celma, the first part of the thesis would not have been possible without his constant help. In addition, I also thank Mohamed Sordo for his collaboration in that part.

I thank as well Perfecto Herrera for his useful advices, Emilia Gomez and the weekly meetings 'Aula 316 Plus', which helped me to gain insights on my research work.

I need also to thank Cyril Laurier and his help in the listening experiments, as well as all those colleagues who participated on it; specially thanks to Jordi Janer and Esteban Maestre for his help on the speech recognition part; Pedro Cano for his initial feedback and support; Hendrik Purwins for his feedback and interesting discussion in the presentations; the freesound team, mainly Bram de Jong and Gerard Roma, who were always disposed to answer my technical questions; and the technical team as well, specially Koppi; and those who read some of my manuscripts, Justin Salamon, Graham Colleman and Pablo Garcia. I also thank Gillet and Richard for allowing me to use their ENST drum database. And those who discussed and gave me feedback when I required it, Martin Haro, Martin Blech, Ferdi Fuhrmann and Stefan Kersten.

Thanks to my family and my colleagues and friends at the MTG, with whom I have been sharing more than knowledge and have contributed to be an unforgettable long year. Finally, I would specially like to thank Paqui and Marcos for their support.

Contents

Chapter 1

Introduction

Profound changes in the way masses are using technology are contributing to the creation and development of new models of production, which rather than being based on hierarchy and control, they are based on community and spontaneous collaboration of crowds. This is giving rise to the emergence of online communities that serve as platform for large repositories, created around these web sites, where users share high volumes of multimedia content. Developers of these new sites may provide the required tools to ensure their success, based on four basic principles: openness, peering, sharing and acting globally, (Wikinomics, [54]).

This thesis presents the framefork of the Semantic Web where online communities are currently growing rapidly. Some of their main characteristics and problems are described as well, selecting as main focus for the thesis the retrieval of sound effects in large collaborative databases. Two different issues have been addressed: the annotation of these large databases by means of collaborative tagging, and a study in alternative sound interfaces based on sound search by phonetic similarity.

Experiments have been performed in *Freesound.org*, a collaborative database where users share and browse sounds by means of tags, and content–based audio similarity search. A short description of it can be found in the following sections. Section 1.1 presents an introduction to the annotation of such large databases, followed by a short description of *Freesound.org* in Section 1.2. Finally, Section 1.3 describes my personal motivation to carried out this work.

1.1 Annotation of Collaborative Databases

Since 2004, collaborative tagging seems a natural way for annotating objects, in contrast to using predefined taxonomies and controlled vocabularies. Internet sites with a strong social component (e.g. *last.fm*, *flickr*, and *del.icio.us*), allow users to tag web objects according to their own criteria. The tagging process can improve then, content organization, navigation, search and retrieval tasks [23].

Audiovisual assets can be manually and automatically described. On the one hand, users can organize their music collection using personal tags like: *late night*, *while driving*, *love*. On the other hand, content–based (CB) audio annotation can propose, with some confidence degree, audio related tags such as: *pop*, *acoustic guitar*, or *female voice*. It is clear that both approaches create a rich tag cloud representing the actual content. Still, automatic annotation based solely on CB cannot bridge the Semantic Gap. Hybrid approaches, exploiting both the wisdom of crowds and automatic content description, are needed in order to close the gap.

In this sense, *Freesound.org*, a collaborative sound database, contains both elements: it allows users to annotate and retrieve sounds based on tags and community data, and it also allows to retrieve similar sounds to a given one, according to audio similarity. A presentation of *Freesound.org* can be found in the following lines.

1.2 Introduction to Freesound

Freesound.org[1] is a collaborative sound database where people from different disciplines share recorded sounds and samples under the Creative Commons license, since 2005. The initial goal was to giving support to sound researchers, who often have trouble finding large sound databases to test their algorithms.

After three years since its inception, *Freesound.org* serves around 20,000 unique visits per day. Also, there is an engaged community—with more than half million registered users—already accessing 54,000 uploaded sounds.

Users can navigate through the database and search sounds by means of tags, full text descriptions, usernames or even filenames. When a user finds an interesting sound, he/she can download it, add comments to it, rate the sound, or even notify if it is an abusive or illegal sample. When a user uploads a new sound to *Freesound.org*, she is asked to tag it and describe it in order to make the sound searchable. Furthermore, by means of collaborative

[1] http://freesound.org/

tagging, other users can extend the sound annotation. There is also the possibility to geotag a sample, or add it as part of a sample pack. Moreover, in the forum users can ask or answer other users questions, or interact with the community sharing different kinds of interests.

1.3 Personal Motivation

If I should summarize in just two words my personal motivation on this work, these would be 'collaboration' and 'sound', which obviously includes music. Moreover, research in topics that many people can benefit from it are also a great motivation, combining it with my passion towards 'music', which has accompanied me since I was a child. These were my first main motivations when I joined the Freesound Project. Carrying out this research within it has made me understand (and develop) some of the principles that are driving to imaginative new ways of innovation and success, conducting to a more critical culture in a more equitable community of better-informed people. In fact, this thesis has also been the result of constant collaboration of different brilliant people, who have been motivating me for working very hard during this year, being part of the Music Technology Group, a great place to learn and touch a bit of the talent of its members.

Chapter 2

Definition of the Problem

Technological advances in computer hardware and software, networking technologies and the move to the Internet as platform for collective multimedia creativity have caused, among others, the proliferation of online communities, where millions of people are connected sharing information and interests. The latest set of applications have transformed the Web from a mere document collection into a social space [38].

All this is happening under the new developments of the Web 2.0, which comes from the ideas of the Semantic Web and seen as an extension of the first Web 1.0. A short description of this evolution is presented in Section 2.1.

Users are transforming the Web and placing new demands on businesses daily. From open source to open content, new forms of organisation, production and distribution are emerging [4]. In the present work, we are interested in multimedia databases that have been created collaboratively around online communities. Investigating the complex social-technological interactions of these kind of systems is 'paramount to our ability of designing intelligent information systems that can take advantage us through of the new online universe' [38]. With this purpose, a summary of the main characteristics and their problems is presented in Section 2.2.

Finally, from all the points examined in Section 2.2, we focus on the retrieval issues of sound effects in large collaborative databases for the present work (Section 2.3). Hence, two different aspects are covered. First, the annotation of such databases by means of collaborative tagging, where the *Freesound.org* Collection has been chosen for the experiments. Second, the study of novel techniques for retrieving sounds effects. Large collections have different levels of description, so there is a need of searching sounds at different levels of criteria. The proposed approach gives the user the possibility of retrieving sounds by means of phonetic similarity and thus enriching or complementing the keyword-based search of such large databases.

2.1 The Scope of the Semantic Web

The original architect and visionary of the **Semantic Web**, Tim Berners-Lee, defined it in 1998 as 'the Web of data (and connections) with meaning in the sense that a computer program can learn enough about what data means to process it' [5], using similar processes to human deductive reasoning. Then, Semantic Web Technologies should organise and find information based on meaning, not just text, for instance, systems that understand where words and phrases are equivalent [26]. These technologies aim to provide formal description of concepts needed to integrating information from heterogeneous sources. For instance, the XML project (eXtensible Markup Language) was initiated to provide an extensible, machine-interpretable language for storing, communicating, and interpreting information [42]. In combination to XML, RDF (Resource Description Format), a language for defining metadata vocabularies, has been largely used for enabling metadata, interoperability and compatibility. Furthermore, Web Ontology Language (OWL) is also an important element of the Semantic Web. It also aims on giving formal representations of concepts, terms and relationships within a given knowledge domain. Reasoning by computers is allowed thanks to this formal conceptualisation [19]. Hence, content may manifest as descriptive data stored in Web-Accessible Databases.

The extension of the Sematic Web where users read and write content in an active way has been called 'the participatory web', 'Social Semantic Web' or better known as **Web 2.0**. This term became notable after the first O'Reilly Media Web 2.0 Conference in 2004. It aims on enhancing creativity, information sharing, and, mainly, collaboration among users where they not only surf the web, but work the web [1].

Some of the features of the Web 2.0 are related to latest user-improved versions of the World Wide Web, open source productions and 'end-to-end' architectures. All these allow simultaneous and unco-ordinated innovation, a basic tenet that has ensured a bottom-up, user-driven innovation [4].

Millions of people meet online to enjoy sharing information, conducting online research or participating in online network communities. This have led to the development and evolution of web-based communities and hosted services, such as social-networking sites, multimedia sharing sites, wikis, blogs, and folksonomies. Despite there is no accepted definition for the term **online community** [45], it was defined by Howard Rheingold, an early pioneer, as 'cultural aggregations that emerge when enough people bump into each other often enough in cyberspace' [48].

Members of these communities have a shared goal, interest or activity that provides the primary reason for belonging to it [63]. Moreover, members have often access to shared resources, such as multimedia content. This is the case of Freesound, where its community users are around a collaborative sound database. The following section summarizes some of the main characteristics of these kind of collaborative databases in online communities.

2.2 Collaborative databases and Online Communities

The recent proliferation of social communities that are around collaborative multimedia databases, such as Flickr, Youtube or Freesound, have provoked the demand for guidelines on how to build successful online communities. In this section I have summarized some of the main characteristics and problems that these communities can have, emphasizing the social and the retrieval aspects. These are community-based, retrieval, content, intellectual property (IP) and digital rights management and finally technical infrastructure issues.

Community-based issues

User communities are created around the repositories. In order to support user communication, blogs, forums and newsletters can be very useful. **Cultural and social aspects** are also important. Cultural barriers, such as the language of the interface or specific content generated by certain users can be problematic sometimes. In addition, **user information** is relevant when understanding how the system should be changed aiming on improving retrieval, recommendation, ontologies, etc.

Besides all that, 'instead of investing in expensive copyright protection systems, information may be commodified through tried-and-tested methods, but this is only possible if due respect is paid to the creative and generative social character of an active user community' [4].

Retrieval issues

Effective retrieval systems are essential in large multimedia databases in order to find elements easily. As stated in [37], good retrieval begins with a **sound organization scheme** for the data, but also depends on sophisticated **user interfaces**. These ones may be designed to help the user understand the data categorization and to meet his/her interests. Moreover, user **usability** and **evaluation** are important issues to take in consideration when designing interfaces [25], as well as knowledge about the communities [11], such as the motivation, interests and behavior of the main body of users, in order to support creation, management and reuse of knowledge in a social context.

Large multimedia collections have different levels of description. Thus, the content retrieval should consider different criteria. The management of digital media needs **semantic descriptors** aiming at making metadata easily searchable and reusable to support possible new users (including computers) and applications [6]. Some sound categorization schemes have been proposed in the literature, but still, none of the existing ones has been positioned as main reference. In this sense, the MPEG-7 can be mentioned, since it offers a framework for the description of multimedia documents [33]. Other sound categorization schemes selected to present here are the ones proposed by Schaeffer and Gaver. The approach proposed by the first one was focused on attributes of the sound itself, without any reference to the source causing it (e.g *pitchiness*, *brightness*), [41]. Gaver [13] introduced a taxonomy of sounds, on the assertion that they are produced by means of interaction of different materials.

Visualization such as the FOAF project (Friend-of-a-friend) is also selected to mention here. The FOAF develops ways to describe both the properties of people (date of birth, age, nickname) and social relationships (interests, places of employment, group affiliations). The most interesting think from a social viewpoint is the fact that one person 'knows' another. In the application of the FOAF relations studied in [42], the authors saw a very rich structuring of the data that was not readily captured in ontologies and yet which was closer to the meaning of social life of the online community they studied.

Since users of the communities created around collaborative databases are both producers and consumers, **collaborative tagging** plays a crucial role concerning the annotation of the multimedia content and thus its retrieval. It can be inferred that active users are influential in such annotation as well.

Retrieval of multimedia content in large databases is not a trivial task, since sometimes it can be a more complex problem. For example, in the case of Freesound, if the retrieval is not about concrete sounds, but group of sounds. Sophisticated navigation systems should then be designed for such purposes aiming at facilitate collective music creativity.

Finally, the effectiveness of the retrieval can be improved by means of **recommendation systems** [47], such as **collaborative filtering**, (when one person searches and he/she is provided with information which has been useful to other people previously [17]), or even with **collaborative retrieval** [3] for some Computer Supported Cooperative Work (CSCW) applications, to the case when many people search for the same information.

Content issues

Assessing the **content quality** and the **authority of contributor** are relevant issues. For example, in Wikipedia[1], an article is considered to be a good one if its author is an authority in the subject area that the article belongs to. In addition, the **peer review** model considers that a review from an authority is valuable. Hence, by means of changing and making corrections in the article, the expertise of reviewers improves the quality and serves to show consensus among other users as well. However, there is some criticism concerning reliability and content accuracy, as well as for its susceptibility to vandalism, but the active Wikipedia community rapidly and effectively repairs most damage [59].

In the case of Freesound, we can differentiate two main parts at this point: sounds and metadata. Hence, **content-based** techniques can be applied in order to measure some aspects concerning the quality of the sounds, such as duration, bps, etc. **rating system** is also applied as a kind of quality measure, but it is not effective when there is a lot of material unrated.

Moreover, **input control** is important to ensure the license of the content when it is required. Each sound of the Freesound Collection has been passed through a moderation stage that guarantees the Creative Commons License.

Since repositories are dynamic, content modification, update or removal are also important issues in order to guarantee the usefulness of such databases. How up to date and the ability to work collaboratively on the improvement are relevant as well.

Intellectual property (IP) and digital rights management

Many of these communities choose **Creative Commons attributions** for the **authoring and creation** of their multimedia content. 'Creative commons thinking will have an impact on any cultural organisation at a number of different levels. First, it brings its own culture to challenge yours. (...) A new culture is emerging that is reshaping not only the 'who' but also the 'how' of cultural production, so that it is transnational, more egalitarian, more transparent to its users, less deferential, much more diverse, and above all, self-authored. (...) Cultural commons thinking prepares us for a mutual world in which we will all be authors, publishers and real-time reporters' [4].

Economical sustainability is an important aspect to consider, such as business models, e-commerce systems, etc. for the creation and maintenance

[1] http://wikipedia.org

of effective repositories in a mid-term sense. This can be achieved, for instance, with models honoring the intellectual property rights of contributors aiming on developing the resources needed to maintain the usefulness and freshness of the repositories and their technical quality [37].

Technical infrastructure issues

Appropriate technical infrastructure is needed for the correct operation of large on-line databases, for instance, security of the system, telecommunication capabilities, updates, operating and maintenance staff, etc. **Data management specifications** are also important and they should be covered by the industry standards. In order to publish or get data from the repositories, interoperability between semantic metadata standards (MPEG-7, MPEG-21, XML, XOL, RDF, etc.) is demanded. In addition, **metadata extraction** and **ontology management** can be time-consuming and require the services of both experts in ontology engineering and domain of interest. Moreover, as usage changes, knowledge also changes, and there is thus an evolution of ontologies to take in consideration[26].

Regarding technical capabilities such as bandwidth or quality, different users may have different needs. Finally, the scalability of such databases is important for the guarantee of effectiveness at higher volumes of content, users, transactions, retrieval, etc. In addition, mirrowing approaches can be a solution for that.

2.3 Issues on Retrieval of SFX Selected for the Thesis

From all the points previously examined in Section 2.2, the retrieval issues have been the focus for the present work. Specifically, those related to the retrieval of sound effects in collaborative databases. In these repositories, high volumes of content are generated and annotated in an heterogeneous way. Since computers are still not sufficiently effective interpreting human descriptions or annotations of this multimedia content, the effective management of it is crucial for a correct retrieval and scalability. This is basically the attempt of the first part of the thesis: study the strengths and weakness of collaborative tagging and propose methods to enhance the retrieval of audio content scarcely annotated. The second part of the thesis is in the context of alternative interfaces. We believe that by means of phonetic similarity, the keyword-based search can be complemented, allowing at the same time a more natural retrieval.

2.3.1 Collaborative Tagging in the Freesound Collection

As mentioned in Section 1.2, *Freesound.org* is a large collaborative sound database accessible on the web. Since its inception, it has become very popular within the audio community, for instance, being well-known for its wide range of sounds. Due to its collaborative nature, it is an excellent tool for the present study.

When analyzing this community, various problems have been detected. Some of them derived by the fact that crowds are annotating multimedia content without constraints. Computers should understand and interpret human annotations in an intelligent way. But they are still far from that, and thus retrieval tasks can be sometimes hard.

Related work can be found in Chapter 3, while Section 3.1 gives the framework for semantic audio descriptions and Section 3.2 presents recent work on Collaborative Tagging. Then, Chapter 4 analyzes tag characteristics in *Freesound.org*, which has been chosen for the experiments presented in section 4.2. However, after this study we have detected that there are some sounds which are scarcely annotated, thus frustrating their retrieval using keyword–based search. The main goal of the experiments are thus to enhance semantic annotations in the *Freesound.org* sound collection, and to evaluate whether using content–based audio similarity we can extend sound annotations. Section 4.2 presents the approach used to ”autotag” sounds based on the tags available in their most similar sounds. Finally section 4.3 presents the results and main findings, and section 6.1 draws conclusions and outline future work.

2.3.2 Sound Search by Phonetic Similarity

Large audio collections have many different levels of descriptions. Hence, there is a need to search sounds with different criteria. Specifically, in the second part of the thesis, we address the case of sound search by phonetic similarity. Onomatopoeias can be considered in the middle way between words and sounds, since they try to convey a sound impression with words or simply by phoneme concatenation, evoking thus the source producing it. Furthermore, the use of onomatopoeia is considered a close approach to signal level properties that a wide range of applications could take profit, such as sound understanding, transcription, classification or retrieval. For instance, when the application requires an exhaustive level of specification. Hence, automatic techniques extracting information at the phonetic level could be then useful.

Chapter 3, specifically Section 3.3, presents existing related work in the context of phonetic sound similarity, while Chapter 5 presents the research conducted for the present thesis. Section 5.1.1 shows some problems found when investigating how subjects understand sound by means of onomatopoeia. Then, the proposed system to automatically extract sound information at the phoneme level is presented in 5.2. The different Experiments and obtained results are then described in 5.3. Finally, conclusions and future work for this second part of the thesis is presented in 6.2.

Chapter 3

Research Context and Related Work

Due to the proliferation of multimedia content on Internet and its availability through online repositories, novel ways to annotate, index and retrieve it are crucial for guaranteeing a satisfactory user interaction. By multimedia content we understand text, such as blogs or wikis, images, video and audio. Audiovisual assets can be manually and automatically described, and both approaches can benefit from each other. In this Chapter 3 we present some of the related work in the audio and music domain. First, Section 3.1 introduces existing related work in semantic audio descriptions, giving thus a general research context. Then, Section 3.2 presents the limitations and drawbacks of collaborative tagging. Finally, Section 3.3 presents the specific case of phonetic sound description applied to audio understanding, annotation, transcription and retrieval.

3.1 Automatic Extraction of Semantic Descriptions in the Audio and Music Domain

In this section some related work that deals with automatically enhancing semantic descriptions in the audio and music domain is presented.

In [58], the authors propose a query–by–semantic audio information retrieval system. The proposed system can learn the relationships between acoustic information and words (tags) from a manually annotated audio collection. The learning task is based on a supervised multiclass labeling model, with a multinomial distribution of words over a predefined vocabulary.

Torres et. al propose a method to construct a musically meaningful vocabulary [57]. By means of acoustic correlation using sparse canonical component analysis (sparse CCA), they can remove from the vocabulary those noisy words (not related with the actual audio content) that have been inconsistently used by human annotators.

The *bag–of–frames* (BOF) approach has been extensively used to describe timbrical properties of an audio signal. This approach consists on using Gaussian Mixture Models (GMM) of Mel–Frequency Cepstrum Coefficients (MFCCs), in order to model local spectral features from signals. The approach is used to extract mid–level descriptions from music signals, such as their genre or instrument, but it is also used to perform timbre similarity between songs. In [2], the authors find out that this approach tends to generate false positives songs which are irrelevantly close to many other songs. These songs are called hubs, and the authors propose measures to quantify the "hubness" of a given song. This property affects any system that uses timbrical features to compute content–based audio similarity.

Cano has studied the strengths and limitations of audio fingerprinting, and suggests that it can be extended to allow content–based similarity search, such as finding similar sounds using query–by–example [6]. Moreover, the author proposes a general sound effects classifier, capable of generating verbose descriptions which combine both low–level and high–level semantic approaches (exploiting the synonyms information available from the Wordnet lexical dictionary).

Similarly to the present approach (see 4.2), in [50] the authors propose a non–parametric strategy for automatically tagging songs, using content–based audio similarity to propagate tags from annotated songs to similar, non–annotated, songs.

In [10], the authors present a method to recommend tags to unlabeled songs. Automatic tags are computed by means of a set of boosted classifiers (Adaboost), in order to provide tags to tracks poorly (or not) annotated. This method allows music recommenders to include in a playlist unheard music that otherwise would be missed, enhancing the novelty component of the recommendations.

Last but not least, social tags are a promising way of exploring a music collection [31]. Levy et. al exploit this idea using Correspondence Analysis to visualise an effective low–dimensional semantic space defined by the tags. This allows one, for instance, browsing the collection by moods.

3.2 Collaborative Tagging

Collaborative tagging can be seen as "feral hypertext" [61], as they are *out of control*. Users are not constrained by a controlled vocabulary when annotating the content. According to Walker, one of the most interesting aspects of collaborative tagging is that the whole community benefits from sharing information [61]. However, as stated in [18], "collective tagging has also the potential to aggravate the problems associated with the fuzziness of linguistic and cognitive boundaries". Users' contributions produce a huge classification system that consists in an idiosyncratically personal categorization. The main problems concerning collaborative tagging are: polysemy, synonymy, the basic level variation of the annotations, and data scarcity. Furthermore, spelling errors, plurals and parts of speech also clearly affect a tagging system.

Sometimes, polysemous tags can return undesireable results. For example, in a music collection if one is searching using the tag *love*, the results can contain both love songs, and songs that users like it very much (i.e. a user that loves a *death metal Swedish* song, not related with the love theme).

Tag synonymy is also an interesting problem. Even though it enriches the vocabulary, it presents also inconsistencies among the terms used in the annotation process. For example, *bass drum* sounds can be annotated with the *kick drum* tag; but these sounds will not be returned when searching for *bass drum*. To avoid this problem, sometimes users tend to add redundant tags to facilitate the retrieval (e.g. using *synth*, *synthesis*, and *synthetic* for a given sound excerpt). Yet, there are some approaches to measure semantic relatedness between tags [8]. These metrics could be used to decrease the size of the vocabulary, and also for (automatic) query expansion to increase the recall in the sound retrieval task.

Besides, there is a systematic variation across subjects concerning what constitutes a *basic level* of an annotation. Basic level is related to the way humans interact with the items to annotate. Categories are usually not well defined, and their boundaries are vague [29]. Thus, users tag an item with different levels of detail; experts in a concrete domain tend to annotate with a level of greater specificity than those without domain expertise. For example, a tagger may tag a guitar sound using *guitar, chord* whilst another user may tag it with *electric-guitar, C chord, fender-stratocaster, 1957*. Finally, the scarcity and inequality nature of a collaborative annotation process—where usually a few sounds are well annotated, and the rest contain very few tags—limits the coverage retrieval of a collection.

3.3 Phonetic Sound Similarity for Sound Retrieval

Large audio collections have many different levels of descriptions. Hence, there is a need to search sounds with different criteria. Moreover, there is a wide range of applications which can take profit of phonetic sound similarity, such as sound understanding, transcription, classification or retrieval. Usually, the level of specification required by certain users or applications is very exhaustive. For example, a woman runing with high-heels on wet sand and that sounds like 'plof plof' could be needed in a post-production movie.

This Section 3.3 deals with the last part of the example, some related work on phonetic sound similarity, and its application to sound retrieval by means of onomatopoeia, understood here as combination of phonemes or even syllables. First of all, we are going to present some studies validating the use of onomatopoeic representations of sound for sound annotation, transcription or retrieval. It is worth to say that most of these studies are dealing with Japanese onomatopoeia and are few.

3.3.1 Acoustic-Speech Correlations in Non-Speech Signals

Some studies attempt to take advantage of the semantic relations existing within language in order to build novel techniques for retrieval. For instance, Cano used in [6] semantic networks that complement content-based classification, using the semantic dictionary WordNet[1], which instead of organizing words alphabetically, as traditional dictionaries do, WordNet organizes concepts with several links between them. His system is a promising approach towards high-level descriptors which are closer to human sound understanding.

However, these techniques are still far from acoustic level properties that directly affect source perception, as stated in [51], where the authors used onomatopoeia as closer approach to signal level properties. The system consisted on mapping the acoustic and semantic sound space by the following approach. They represented onomatopoeia in a "meaning space" by means of an inter-word distance metric. Then, they tagged some audio sounds according to the onomatopoeia. Finally, they could observe that the audio within the same cluster shared both semantical and acoustic properties.

In [12] have determined some relationships between phonetic features of onomatopoeia and tonal features of the auditory imagery impressions that

[1] http://wordnet.princeton.edu/

subjects present in their study. For instance, the vowel /i/ was associated with sharp impressions and vowels /u/ and /o/ were associated with dull impressions. The obtained tendencies confirmed that some onomatopoeic features reflected particular impressions of auditory imagery.

The same authors demonstrated in [35] the validity of onomatopoeic representations for identifying acoustic properties of sound by using a classification system based on the similarities of phonetic features using hierarchical cluster analysis.

Our understanding of sound is a subjective and abstract concept, that some studies have investigated in order to find common ways to describe sounds. After the study carried out in [60], it was realised that human sound representation use three basic types of verbal description:

- sound itself by means of onomatopoeia (or combination of syllables), which describe the acoustical sound properties,
- sound situation which describe the context (what, where, when, and how etc. produce it),
- the sound impression that produces on the listener, described usually with an adjective, depending essentially on the subjectivity of each person (e.g if a sound is pleasant, annoying, etc.).

They pointed out that sound situation was the most frequently used type, although 'sound itself' was used for the first description. Then, they proposed a method for retrieving sounds by Japanese onomatopoeia, sound source and adjective. In addition, they reported that retrieval was fast when user had concrete knowledge of the sound. On the contrary, the the use of onomatopoeia was useful in such cases where the user had vague ideas of the sound and furthermore could not specify the source sound.

3.3.2 Onomatopoeia and Environmental Sounds

Onomatopoeia has been considered very effective for situational communication as well as for environmental sounds ([7], [53], [24]).

For instance, Silva et al. [7] presented a preliminary study on acoustic analysis of environmental sounds based on Gaver's sound classification (Gaver, [13]). Similarities between sounds produced by objects with the same type of material and interaction were found. Moreover, the onomatopoeia representations proposed by the studied subjects shared similar phonetic features.

Besides that, Tanaka proposed the use of onomatopoeia to detect machine errors [53].

In addition, the fact of imitating sounds by phonetic similarity or onomatopoeia was called 'sound-imitation words' in [24]. They stated that the main problem in automatic sound-imitation words is the ambiguity in determining phonemes, since it is very subject-dependant. Moreover, they designed a system based on Hidden Markov Models (HMM, [46]) with a phoneme-group set for environmental sounds to automatically transcribe them into Japanese onomatopoeia.

Onomatopoeia words are also a means of symbolic grounding, since sounds are transformed into symbolic representations [24]. The same authors stated that, 'in digital archives, onomatopoeia words may be used for annotations, such as in MPEG-7 [33] is used for sound signals'. This study is specially interesting since they attempt to automatically transform environmental sounds into Japanese words, very similar to the present study. Moreover, they also stated that 'Automatic speech recognition (ASR) systems [27] fail to recognize non-speech sounds, in particular environmental sounds such as friction, impact and electronic sounds'.

3.3.3 Drum Sound Recognition

Here some related work in Percussive Sound Recognition is introduced, since it serves as context for the following Subsection 3.3.4. Many of the works in music retrieval have been focused on melody or query by example, such as query by humming based on melody or query by rhythm. However, in the context of percussive sounds, melody is hardly present, different approaches should be thus considered. Many studies in the literature have attempted to classify drum sounds. P. Herrera [22] found a useful list of twenty features, achieving classification rates above 90% within specific taxonomies. Tindale published a survey of related work in beat detection and drum identification [55]. He also investigated in [56] the classification of snare drum timbres produced by different playing techniques, presenting a system that could recognize subtle differences in timbre, which they believed this was the first step towards a comprehensive system able to transcribe music and providing information at the timbral level.

3.3.4 Voice Percussion Recognition

Reviewing the literature, different works have validated the use of voice percussion recognition as promising approaches for music transcription or composition ([40], [28], [20]). Drum sound recognition (see 3.3.3) looks for acoustic properties that are characteristic of the instrument, but in our case, mapping between the input voice and target instrument sound is only indirect and metaphoric. Query-by-humming foucuses on pitch detection and melodic feature extraction, but these have less relevance in voice percussion recognition, which is primarily concerned with classification of timbre and identification of articulation methods. Differences among individuals in both vocal characteristics and the kinds of verbal expression used add further complication to the task. Examples of these approaches can be found in [36] where their system allows the user to retrieve sounds given their drum pattern by means of voice percussion (beatboxing). Hence, the user search for a sound by natural sounds such as 'ta pum pum ta'.

In [40] they used voice percussion recognition techniques for retrieving drum patterns by means of classification of timbre and identification of articualtion methods. This is an indirect and metaphoric approach, since as they reported, 'voice percussion is not a direct, faithful reproduction of the acoustic properties of actual drum sounds'. So, they exploited the concept of onomatopoeia by mimicking drum sounds by voice, in order to transcribe them to their corresponding phonemic sequences by means of classical speech recognition techniques. Then, they applied sequence matching for retrieving the drum pattern with the highest likelihood score for the given user query. The target drum patterns consisted on four beats of Bass Drum and Snare Drum sounds with no simultaneous beats.

They obtained a recognition rate of about 70%. Finally, they invited for further investigation to different languages and background cultures, since one of the limitations of this work was the data used, which was only obtained from Japanese speakers. Another limitation of this work was that they just considered Bass and Snare drums, inviting to extend it for more percussive instruments, as well as, more complex drum patterns.

Example of a Timbre-based Musical Instrument: the North Indian Tabla

Pitch is the primary basis for sound categories in music, intervals and chords, while timbre is for speech, vowels and consonants. Patel explains in [43] why timbre contrasts are rarely the basis for musical sound systems, with some excepcions such as the North Indian Tabla. Compairing music and speech, two aspects emerge: temporal and spectral profile of a sound. While the first one refers to the temporal evolution of the amplitude, spectral proile refers to the frequency distribution, as well as their relative amplitudes. The discussion is why timbre rarely serves as the basis for musical sound systems. Physical reasons on the way that instruments are excited can answer partly this question; dramatic changes in timbre are difficult or impossible to handle. Cognitive reasons are also important; there are no such timbre intervals, ordered in terms of perceptual distances as in pitch.

However, there is a musical system certainly based on timbre: the North Indian Tabla. The author in [44] studied the perceptual and acoustic resemblance between each stroke of the instrument and its associated vocable. Empirical research was conducted to confirm this link, which is interesting since drum and voice signals produce sounds in a very different way [43], starting from the hypothesis that vocables are a case of sound symbolism or onomatopoeia. So, the point was how timbral differences between drum sounds were mapped onto linguistics. They demonstrated strong connections between the musical and linguistic timbres of the North Indian culture, and they encourage further investigations on similar empirical studies in other cultures with rich percussive traditions, such as Africa and the african diaspora. Speciffically, they found correlations between acoustic properties of drum sounds and the phonetic component of vocables, such as spectral centroid, rate of amplitude envelope decay, duration between the releases of consonanats, fundamental frequency and the aspiration on the balance of low vs. high frequency energy in a vowel.

We find this study interesting since gives evidence of sound symbolism in tabla, and this opens thus new ways of research in other instruments or application contexts. One could think that this is culture dependent, and difficult to generalyze, and in fact it is. However, they performed a perceptual experiment with subjects who were non familiar with tabla drums. The results were quite promising, since in most of the cases, these naive listeners were able to match each vocable with the corresponding tabla sound.

Tabla vs. Drum North Indian Tabla has a well-defined set of onomatopoeia, where each stroke of this instrument is mapped in a commonly accepted set of vocables. This is not the case for drum sounds, where there is no such commonly accepted set of onomatopoeia for describing the different sounds. In [14], Guillet and Richard pointed out that this could be explained by the important role that oral tradition plays in non-western music cultures, where the notation does not play an important role as in western popular music does.

The same authors studied different classification approaches on a simplified taxonomy of drum loops signals. The best result of 89.9% correct recognition rate was achieved by a novel approach of SVM with context. Since each drum loop signals exhibit temporal structure, their approach was based on a sequence model or language model by analogy with large vocabulary speech recognition systems. It consisted on modeling each stroke probability with the probabilistic output of the SVM classifier, as well as adding context information coming from the previous stroke states, modelled with a language model (the transition probabilities were estimated by counting occurrences in each N-gram in the training database). The queries were formulated by means of spoken onomatopoeia and the system compaired them to the rhythmic sequences in the database. They imposed a set of onomatopoeia to the user, validated from a perception experiment performed in [15], (it would be interesting to know the native languages of the participants, since onomatopoeia is language-dependant). This approach is limited to a speaker-dependent recognition and should be extended for speaker-independent recognition.

Chapter 4

The Freesound Collection

This chapter presents the general characteristics of *Freesound.org*, the online community where users share and browse audio files by means of tags, and content–based audio similarity search already presented in previous chapters.

We performed two analyses of the sound collection. The first one (Section 4.1) is related to how users tag sounds. Some well–known problems that occur in collaborative tagging systems were detected (i.e. polysemy, synonymy, and the scarcity of the existing annotations). Moreover, we notice that more than 11% of the collection was scarcely annotated with only one or two tags, thus frustrating the retrieval task. In this sense, the second analysis (Section 4.2) focuses on enhancing the semantic annotations of these sounds. Then, the results are presented in Section 4.3.

4.1 General Characteristics

In this section general characteristics of *Freesound.org* are introduced, as well as the tag behavior in Subsection 4.1.1 and the different tag categories present in the collection, Subsection 4.1.2.

Figure 4.1 depicts, in a log–log scale, all the sounds uploaded by the users. The horizontal axis contains the 1,843 users that have uploaded one or more sounds, ranked by the total number of sounds uploaded. The y-axis shows the number of sounds added by each user. The shape of the curve follows a power–law ($x^{-0.87}$); a few dozens of users uploaded hundreds of sounds, whilst the rest uploaded just a few. In fact, 80% of the users uploaded less than 20 sounds, and only 5 users uploaded more than one thousand sounds each. It is worth noting that these few users can highly influence the overall sound annotation process.

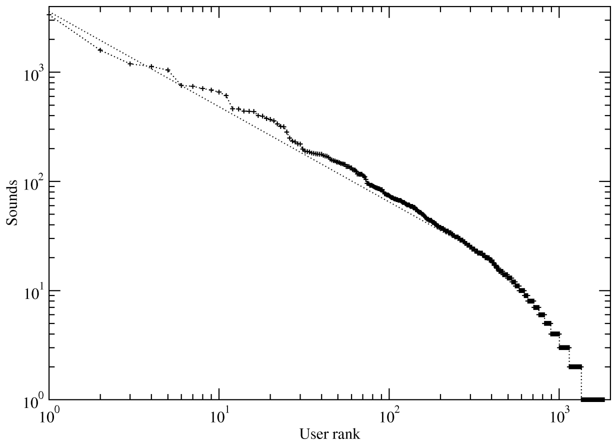

Figure 4.1: A log–log plot depicting all the uploaded sounds by the users. Top–1 user has uploaded more than 3,000 samples. The curve follows a power–law with $x^{-0.87}$, with an exponential decay starting at $x \simeq 400$.

4.1.1 Tag behaviour

In this subsection we provide some insights about the tag behaviour and user activity in the *Freesound.org* community. We are interested in analysing how users tag sounds assets, as well as the concepts used when tagging. The data, collected during February 2008, consists of 44,270 sounds annotated with 12,420 different tags.

Figure 4.2 shows the number of tags used to annotate the audio samples. The x-axis represent the number of tags used per sound. We can see that most of the sounds are annotated using 3–5 tags. Also, around 5,000 sounds are insufficiently annotated using only 1 or 2 tags. These sounds represent the 11% of the total collection. It would be desirable, then, to—automatically—recommend relevant tags to these scarcely annotated sounds, enhancing their descriptions. This is the main goal of the experiments presented in section 4.2.

Interestingly enough, in [6], the author analysed a sound effects database, which was annotated by only one expert. A similar histogram distribution to the one presented in Figure 4.2 was obtained. Specifically, most of the sounds were annotated by the expert using 4 or 5 tags, as it is our case. This could be due to human memory constraints when assigning words to sounds or to any object, in order to describe them [39]. Based on Figure 4.2, we classify the sounds in three categories, according to the number of tags used. Table 4.1 shows the data for each class.

Tag frequency distribution is presented in Figure 4.3. The x-axis refers

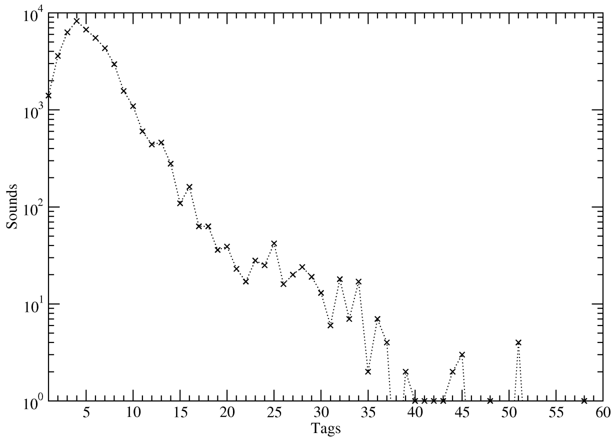

Figure 4.2: A lin–log plot depicting the number of tags per sound. Most of the sounds are annotated using 3–5 tags, and only a few sounds are annotated with more than 40 tags.

Table 4.1: Sound–tag classes and the number of sounds in each category.

	Tags per sound	**Sounds**
Class I	1–2	4,997
Class II	3–8	34,056
Class III	¿ 8	5,217

to the 12,420 tags used, ranked by descending frequency. The curve follows a power–law, with $x^{-0.84}$, with an exponential decay starting at $x \simeq 10^3$. This decay is due to very infrequent or misspelled tags. On the one hand, 44% of the tags were applied only once. This reflects the subjectivity of the tag process. Thus, retrieving these sounds in the heavy tail area is nearly impossible using only tag–based search. To overcome this problem, *Freesound.org* offers a content–based audio similarity search to retrieve audio samples that sound similar to a given one. On the other hand, just 27 tags were used to annotate almost the 70% of the whole collection.

The top–10 most frequent tags are presented in Table 4.2, and it gives an idea about the nature of the sounds available in the *Freesound.org* collection. All these frequent tags are very informative when describing the sounds, in contrast to the photo domain in *flickr.com*, were popular tags are considered too generic to be "useful" [49].

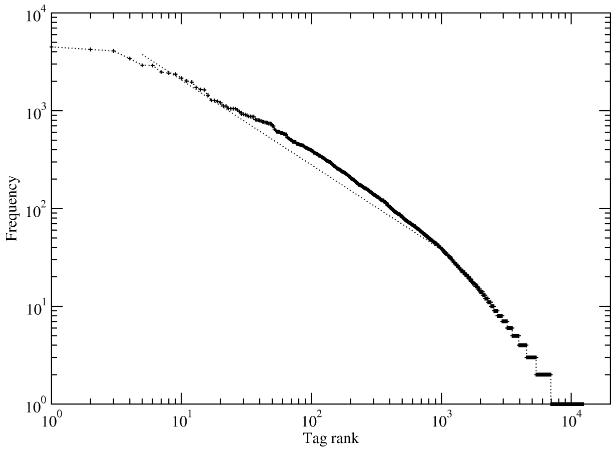

Figure 4.3: A log–log plot showing the tag distribution in *Freesound.org.* The curve follows a power–law with $x^{-0.84}$, with an exponential decay starting at $x \simeq 10^3$.

Table 4.2: Top–10 most frequent tags from Figure 4.3.

Rank	**Tag**	**Frequency**
1	field–recording	4,486
2	loop	4,227
3	noise	4,094
4	electronic	3,417
5	drum	2,916
6	synth	2,898
7	processed	2,482
8	ambient	2,424
9	bass	2,366
10	voice	2,155

4.1.2 Tag categorization

In order to understand the vocabulary that the *Freesound.org* community uses when tagging sounds, we mapped the 12,420 different tags to broad categories (hypernyms) in the Wordnet[1] semantic lexicon. In some cases, a given tag matches multiple entries, so we bound the tag (noun or verb) to the highest ranked category.

The selected Wordnet categories are: *(i)* artefact or object, *(ii)* organism, being, *(iii)* action or event, *(iv)* location, and *(v)* attribute or relation. Figure 4.4 depicts the distribution of tags matched among the Wordnet categories (20.3% of the tags in the collection remain unclassified, and are not presented in the pie chart).

Most of the tags (38%) are related with objects (e.g. *seatbelt*, *printer*, *missile*, *guitar*, *snare*, etc.), or about the qualities and attributes of the objects (30%); such as state attributes (*analog*, *glitch*, *scratch*), or magnitude relation characteristics (*bpm*). Then, some tags (19%) are classified as an action (*hiss*, *laugh*, *glissando*, *scream*, etc.), whilst 11% are related with organisms (*cat*, *brass band*, etc.). Finally, only a few tags were bound to locations (e.g. *iraq*, *vietnam*, *us*, *san francisco*, *avenue*, *pub*, etc.). Therefore, we can conclude that the tags are mostly used to describe the objects that produce the sound, and the characteristics of the sound. In this case, the wisdom of crowds concords with the studies of Schaeffer [41], and Gaver [13]. The first one focused on the attributes of the sound itself without referencing the source causing it (e.g *pitchiness*, *brightness*). While the second one introduced a taxonomy of sounds, on the assertion that they are produced by means of interaction of materials.

4.2 Experiments

This section focuses on enhancing the semantic anotations of the sounds scarcely annotated with only one or two tags. The dataset selected is explained in Subsection 4.2.1. Our goal is to evaluate the quality of the recommended tags, for some specific sounds available in *Freesound.org*. By means of content–based audio similarity, we use a k–NN classifier (Subsection 4.2.2) that, given a sound, selects a set of candidate tags available from the most similar sounds. Then, the evaluation process is based on human assessment in order to evaluate the perceived quality of the candidate tags. The procedure is described in Subsection 4.2.3. Three subjects validated each candidate tag for all the sounds in the test dataset as it is explained in 4.2.4.

[1] `http://wordnet.princeton.edu/`

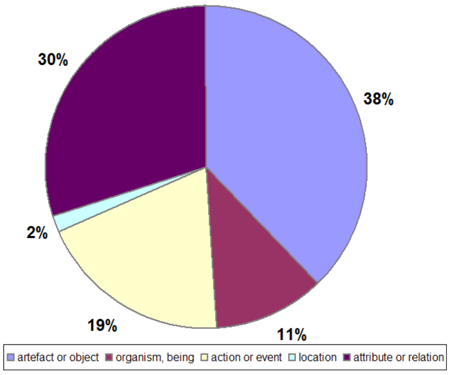

Figure 4.4: Pie chart depicting the distribution in Wordnet categories for around 10,0000 mapped tags.

4.2.1 Dataset

The sounds selected for the experiments were a subset of the Class I (see Table 4.1). We selected those sounds whose tags' frequency was very low (i.e. rare tags, in the ranking of $\sim 10^4$ in Figure 4.3). In fact, all the sounds which were annotated with one tag whose frequency was equal to 1 were selected. Also, for the sounds annotated with 2 tags, we selected those which had at least one tag with frequency 1. The test dataset for the experiments consists of 260 sounds. The goal, then, is to extend the annotation of these sounds, unsufficiently annotated with one or two very rare tags.

4.2.2 Nearest–neighbor classifier

We used a nearest neighbor classifier (k–NN, $k = 10$) to select the tags from the most similar sounds of a given sound. The choice of a memory–based nearest neighbor classifier avoids the design and training of every possible tag. Another advantage of using an NN classifier is that it does not need to be redesigned nor trained whenever a new class of sounds is 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 [6].

The similarity measure used is a normalized Manhattan distance of audio features belonging to three different groups: a first group gathering spectral and temporal descriptors included in the MPEG-7 standard [33]; a second one built on Bark Bands perceptual division of the acoustic spectrum, using 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 [21]. The normalized Manhattan distance of the above enumerated features is:

$$d(x, y) = \sum_{k=1}^{N} \frac{|x_k - y_k|}{(max_k - min_k)} \tag{4.1}$$

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

4.2.3 Procedure

Our technique for calculating the candidate tags consists on finding the *10–th* most similar sounds from the *Freesound.org* database, for a given seed sound of the test dataset. Therefore, given a seed sound, we grab the available tags from the similar sounds. A tag is proposed as a candidate if it appears among the neighbors over a specific threshold. For example, a threshold of 0.3, means that a tag is selected as candidate when it appears at least in 3 sounds of the 10 nearest neighbors. In this manner, we select a set of candidate tags for each sound in the test dataset.

The experiments have been computed using two thresholds: 0.3 and 0.4. When using a threshold of 0.3 the number of candidate tags is, obviously, higher than for 0.4, but also it can be more "noisy" tags, since it is using a less constrained approach. Afterwards, all the candidate tags will be evaluated by human assessment, so we compare the differences among both thresholds in section 4.3.1.

4.2.4 Evaluation

In order to validate the candidate tags per test sound, we use human assessment. It is worth noting that neither Precision nor Recall measures are applicable as the test sound contains only two or less tags, and these are very rare in the vocabulary. The aim is to evaluate the perceived quality of the candidate tags. With this purpose, we performed a listening experiment where the subjects were asked to listen to the sounds, and decide whether they agreed or not with the candidate tags. For each candidate tag, they had to select one of these options: *Agree* (recommend candidate tag), *Disagree* (do not recommend), or *Don't know*. Each sound was rated by three different subjects.

Similar to [58], to evaluate the results we group human responses for each sound s, and score them in order to compact them into a single vector per sound. The length of the vector is the number of candidate tags of s. Each value of the vector, $\mathbf{w}_{s,t_i}$, contains the weight of the subjects' scores for a candidate tag t_i in sound s. If a subject agrees with the candidate tag, the score is $+1$, -1 if disagrees, and 0 if she does not know. The formula for calculating the weight of the candidate tag in s is:

$$\mathbf{w}_{s,t_i} = \frac{\#(PositiveVotes) - \#(NegativeVotes)}{\#Subjects} \tag{4.2}$$

A candidate tag is recommended to the original sound if $\mathbf{w}_{s,t_i}$ is greater than zero, otherwise, the tag is rejected (either because it is a bad recommendation, or the subjects cannot judge the quality of the tag). For example, given a candidate tag t_i for s, if the three subjects scored, respectively, $+1$, -1, $+1$ (two of them agree, and one disagree), the final weight is $\mathbf{w}_{s,t_i} = 1/3$. Since this value is greater than zero, t_i is considered a good tag to be recommended. Also, we use these weight values to compute the confidence agreement among the subjects. First, we consider all the sounds where the system proposed j candidate tags, S_j. We sum, for each sound $s \in S_j$, the weights of all the candidate tags t_i whose values were greater than zero. Then, we divide this value with the total score that the candidate tags would had if all the subjects would agree. The formula for calculating the agreement of S_j sounds, A_j, is:

$$A_j = \frac{\sum_{s \in S_j} \left[\mathbf{w}_{s,t_i} > 0\right]}{\#Subjects \cdot \left[\sum_{s \in S_j} length\,(s)\right]} \tag{4.3}$$

Similarly, to compute the agreement of the bad candidate tags, we use the weights of candidate tags whose values were lesser than zero ($\mathbf{w}_{s,t_i} < 0$), in the numerator of the equation 4.3. Finally, to get the total agreement for all the sounds in the test set, A_{total}, we use the weighted mean of all A_j, according to the number of sounds in A_j.

4.3 Results

4.3.1 Perceived quality of the recommended tags

Using 10–NN and the content–based audio similarity, and setting a threshold of 0.3, the system proposed a total of 781 candidate tags, distributed among the 260 sounds of the test dataset. Besides that, setting a threshold of 0.4 the system proposes 358 candidate tags, which represents almost the half compared with a threshold of 0.3.

Table 4.3: Percentage of recommended tags, with confidence agreement among the subjects. The table shows the results using thresholds 0.3 and 0.4 (in parenthesis, it is shown the total number of candidate tags).

Threshold	**Recommend tag**	%	A_{total}
0.3 (781)	Yes	56.6%	0.74
	No	31.59%	0.62
	Don't know	11.41%	—
0.4 (358)	Yes	66.23%	0.78
	No	23.11%	0.58
	Don't know	10.66%	—

Table 4.3 shows the human assessment results. As expected, a slightly higher percentage of candidate tags were recommended with a threshold of 0.4. Yet, using a threshold of 0.3, more than half of the candidate tags (56.6%) were finally recommended to the original sounds, with an agreement confidence of 0.74. This human agreement is sufficiently high to rely on the perceived quality of the recommended tags. The rest of the candidate tags (43.4%) were not recommended, either because the tags recommended were not appropiated (31.59%), or the tags were not sufficiently informative (11.41%). Even though with a threshold of 0.3 we get less percentage of recommended tags, the absolute number of candidate tags is more than twice the ones with a threshold of 0.4. Therefore, we can consider a threshold of 0.3 as a good choice for this task.

4.3.2 Recommended tags per category

On the one hand, using a threshold of 0.3, we have enhanced the annotation of 200 sounds, which represent the 77% of the sounds in the test dataset used. The rest of the sounds (60) from the test set did not get any plausible tags to extend its current annotation. On the other hand, with the threshold of 0.4 we are able to enhance the annotation of half of the sounds (128 sounds out of 260).

Table 4.4 shows the results using a threshold of 0.3, and it classifies the sounds according to the categories defined in Table 4.1. We can observe the number of sounds per class, after extending the annotation of those 200 sounds. Note that, after applying the recommended tags, most of the sounds have 3 or more tags, and some even have more than 8 tags. However, there are 20 sounds that still remain in Class I. This happens because before the experiment they only had one tag, and now they have another one, the one

Table 4.4: Number of sounds in each category after automatically extending the annotations of 200 sounds from the test dataset.

	Tags per sound	**Sounds**
Class I	1–2	20
Class II	3–8	171
Class III	> 8	9

recommended.

4.3.3 Analysis of the recommended tags

The results obtained so far look promising; using a simple classifier we were able to automatically extend sound annotations that were difficult to retrieve. However, we did not introduce yet the characteristics of the recommended tags. E.g. are they popular tags (in terms of usage)? Or, are the recommended tags too specific for the sound, thus not improving the search and retrieval tasks?

First of all, there are some tags that the subjects could not easily evaluate (i.e. they rate them as *Don't Know*). These tags are related with how the sound was created. They cope with different aspects of the sound creation. E.g. a particular algorithm (*subtractive* synthesis), software products (Native Instruments *reaktor*), or even a vintage synthesizer (*ppg* WAVE synthesizer), probably recorded using its VST plug–in version. It seems reasonable that, in these cases, the subjects could not assess the relevancy (and correctness) of the candidate tags.

Furthermore, due to the classifier method used (k–NN), there is a strong correlation among the more frequently proposed tags, and their frequency of usage (rank position in Figure 4.3). Table 4.5 shows the most frequently candidate tags, including how many times they have been proposed (second column) in the 260 test sounds, and its overall ranking of usage by the *Freesound.org* community (third column). The ten most proposed tags are also in the top–15 ranking of frequency use. Although our approach is prone to popular tags, it allows the users now to get a higher recall of those scarcely annotated sounds, when doing a keyword–based search.

An example of a tag less useful for automatic tagging is *multisample*. Specifically, this tag is very frequent in the *Freesound.org* collection, but it has been used by only 11 very active users. All subjects during the evaluation always marked *multisample* as a bad candidate tag. So, these few users had a negative impact in the tag recommendation process. But this tag has a clear sense in the context of a group of sounds. That is, when the same

Table 4.5: Top–10 candidate tags with the number of times being proposed (second column). Also, frequency usage (tag rank) in *Freesound.org*, is shown in the third column.

Tag	**Frequency**	**Rank**
field–recording	60	1
loop	37	2
drum	27	5
noise	24	3
voice	24	10
bass	21	9
synth	21	6
electronic	18	4
ambient	15	8
percussion	16	15

sound is recorded at different pitches (C4, C5, C6). In this case, all the three sounds are annotated with *multisample*. Obviously, this candidate tag has no meaning to be recommended to a single sound.

Chapter 5

Sound Search by Phonetic Similarity

5.1 Preliminary Assumptions

Onomatopoeia is the creation or use of words that sound like the items or actions they name or refer to. Much onomatopoeia seems to fall into the following categories [64]:

- Mechanical Onomatopoeia. Machine noises such as buzz, beep, whirr, click, clack, clunk, clatter, clink. Notice the group of words that begin with cl.

- Fast Motion Onomatopoeia. Words that convey the sound of speed seem often to begin with the letter s or z. Boing, varoom/vroom, whoosh, swish, swoosh, zap, zing, zip, and zoom are some examples.

- Musical Onomatopoeia. Some are associated with specific music instruments, such as the twang of a guitar, or the plunk for a keyboard. Others imitate a metallic sound, often end in ng: ting, ding, ring, clang, bong, brrring, jingle, and jangle. Then there are some that clearly evoke wind instruments, like blare, honk, and toot; and another group that seem percussive, like rap, tap, boom, rattle, and plunk. A person making music without an instrument might hum or clap or snap.

- Food Preparation and Eating Onomatopoeia. Food may crackle or sizzle and oil may splatter when cooking. Drinks when pouring may go splash, kerplunk, or gush, but hopefully they wont drip, and when we open a soft drink, it will probably fizz. When its time to eat, were likely to nibble, munch, gobble, and crunch.

- Fighting Onomatopoeia. Action words present in comic books during fighting scenes are onomatopoetic: pow, bif, bam, whomp, thump, smash, zowie, bang, and wham are some of them.
- Animal Onomatopoeia. In different parts of the world, words used for animal sounds are quite different. Sheep do not universally go baa, nor do ducks quack everywhere in the world.

Onomatopoeia words exist in every language, although they vary in each. Onomatopoeias usually seem to have a tenuous relationship with the object they describe [62]. Native speakers of a given language may not question the relationship, but because they may differ considerably between languages, non-native speakers might be confused. For instance, a dog goes woof-woof in English, wau-wau in German, ouah-ouah in French, or guau-guau in Spanish. In addition, not only each language has its own onomatopoeias, but also the usage tradition even varies within speakers of a certain language, as it has been seen after the online survey conducted for the present thesis.

5.1.1 Online Survey

An online survey was designed to understand how different people make use of onomatopoeia when describing sounds. Moreover, it also served to verify if there is sufficient onomatopoeia usage tradition. Another important aspect was to see the differences between people from different nacionalities. Participants had to listen to a set of sounds and decide which onomatopoeia corresponded to each sound. Some sounds had more than one onomatopoeia. Hence, more than forty onomatopoeia (selected from the above categories) could be answered in total. Twenty participants from six different nationalities were asked to participate in the survey answering in English, even for most of them this was not their native language.

Analyzing the responses, it was quite evident that an onomatopoeia usage tradition was missing in many of the answers of non-english native speakers. However, this was not the case of english natives, who usually answered using established onomatopoeia words. Evidence of that can be the following examples. Onomatopoeia words evoking movement, such as swoosh or whoosh, seem to have no translation to i.e. Spanish. In such cases where the participant did not know an exact onomatopoeia word, he/she tend to imitate the sounds with phonemes, while english natives annotated with that two onomatopoeia words. Some of the answers from spanish participants for that were fiu, glush, suff, wham or zum. As it can be seen, the results from the Spanish participants to a whoosh were too heterogeneous and difficult to evaluate with a single word-distance or string-matching technique. Similar examples were found in other languages.

Another interesting point was that participants used onomatopoeia as a way not only to symbolize sound timbre, but also rhythm. This is the case of a horse gallop sound, where some answers were capatum, tacatac, tacatan, tocotoc or trocotro. Even the phoneme concatenations are different, the sequences suggest some kind of rhythm pattern similarity.

As summary, we can conclude that it seems not to be a commonly way of onomatopoeia usage (at least in non-english langugages), even between native speakers of the same language. Most participants, specifically non-english natives, tend to imitate timbrically sounds by phoneme concatenation instead of using an onomatopoeia word. Hence, the procedure would be finding phonetic similarities at the phoneme-level, instead of word-level.

5.2 Proposed System

Figure 5.1 shows the proposed system to automatically extract information of the sound timbre by phonetic similarity from non-speech signals.

It is basically an Automatic Speech Recognition Sytem (ASR), and specifically it is based on the open source Julius Speech Recognition system [30]. It is composed by two models: the acoustic and the language model. On one hand, the acoustic model contains the Hidden Markov Model (HMM) and the Viterbi backtracking algorithm (Rabiner, [46]) for each phoneme. In our case, each one has three states: attack, sustain and release. Hence, we are considering 'context': from where and to where is going each phoneme is modeled as well. ASR systems usually use libraries of acoustic models trained with large number of subjects within a certain language. We have selected the Japanese AM library from the open source Julius Speech Recognition system. The models are coded as text in the HTK format[1]. Since we use Spanish phonetic symbols as output of the ASR system, the Japanese acoustic models have been converted by using SAMPA[2].

On the other hand, the language model contains the vocabulary and the grammar to be recognized by the system. The vocabulary can be single phonemes (e.g. /p/, /u/, /m/, /t/, /a/, etc.), syllables (e.g /pum/, /ta/, /tum/, etc) or even words. Thanks to the grammar, we are able to allow loops or concatenation of phonemes and words. We also allow silence between two

[1] Cambridge Universitys Hidden Markov Model Toolkit (HTK). http://htk.eng.cam.ac.uk

[2] Speech Assessment Methods Phonetic Alphabet (SAMPA) is a computer-readable phonetic script based on the International Phonetic Alphabet (IPA).

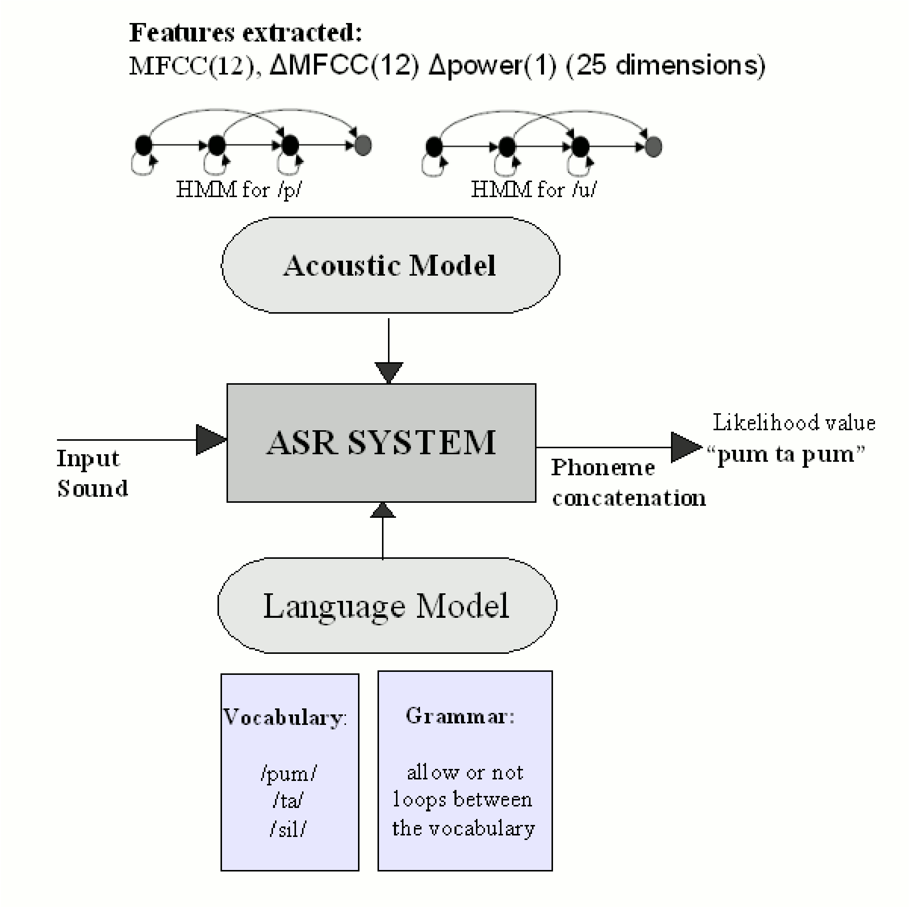

Figure 5.1: Automatic speech recognition system applied to extract information from non-speech sounds at the phoneme level.

consequetive phonemes or syllables when building the Finite State Network (FSN) of each word in the vocabulary. The output state of a certain phoneme has then two possible transitions, either silence or another phoneme.

When a sound is input into the system, vectors of 25 dimensions are calculated for each frame observation. The extracted descriptors per frame are the first twelve Mel-frequency cepstral coefficients (MFCC, [32]), their first moments and the derivative of the Energy Signal. The sampling frequency rate is set to 16 kHz and the window size in number of samples is 400. Classical Viterbi algorithm is then calculated; each node stores the cummulative cost and its preceding node. Since it is backtracking, the best path is found starting from the last frame.

The system output is then a sequence of phonemes (if we analyze it at the phoneme level) and a cost value indicating the cummulative cost above explained when calculating the most likelihood sequence.

5.3 Experiments

5.3.1 First Experiment

The first experiment was conducted to validate the proposed approach for recognizing phoneme sequences from non-speech signals without semantic information.

Dataset The sound database used for this first experiment was a 300 percussive sound database manually built with this purpose (containing bass drums, snare drums, hi hats, cymbals, etc.). All sounds were extracted from the *Freesound.org* Collection.

Language Model It has been built in a way that only single phonemes and concatenation of them were allowed. Hence, the vocabulary is composed of about 31 spanish phonemes transcribed to their respective SAMPA representations.

Results Results showed that the present approach is able to recognize phoneme sequences, suggesting onomatopoeic patterns in some cases. The following table 5.1 shows some of these examples when the input were bass and snare drum sounds respectively.

Table 5.1: This table shows some phoneme outputs for bass and snare drum sound inputs. Note that some phoneme patterns are repeated among different sounds (each row represent different sounds).

Bass Drums	**Snare Drums**
poun	peua
prumn	pua
zlr	pua
zrlamonmn	spuaa
panmfu	duaa
panmiam	rua
cllynmn	ruaj
cpomn	pruaj

Hence, some phoneme sequences were more frequent than others. For instance, from a total of 80 snare drum sounds, the most frequent phonemes were /a/ (frequence of 28%), /u/ (25%), /p/ (23%), /n/ and /m/ (both with 19% of frequence). But still, these phoneme frequencies are not sufficiently high to be able to generalize the results. Moreover, many sounds were described with single phonemes, rather than with onomatopoeic sequences.

5.3.2 Second Experiment

This experiment served to inductively determine which onomatopoeia might be the best for the concrete case of describing snare and bass drum sounds. Gillet and Richard performed in [14] a listening experiment in order to determine the best onomatopoeia to describe drum sounds. Around 30 subjects were asked to select, from a given list of onomatopoeia, the one they considered fit the best for each drum sound. Hence, Table 5.2 shows the results for bass and snare drums. As we can observe, poom [pum], boom [bum] and too [tu], were the most commonly used onomatopoeia for bass drums. Analogy, tcha [tSa], ta [ta], too [tu] and tss [ts] were the frequent ones for snare drums. In addition, these are the onomatopoeia selected for the following experiments.

Besides, they reported that [ta] and [tSa] are often used to denote the mixture of snare drum and bass drum. This could be explained by the fact that subjects tend to link the onomatopoeia to the most salient instrument.

Table 5.2: Results of the listening experiment conducted in [14]: the chosen onomatopoeia are represented in bold which correspond to the most frequent ones.

Sound	Onomatopoeia	Freq.	Onomatoporia	Freq.
Bass Drum	**[pum]**	**36**	**[bum]**	**16**
	[tu]	**17**	Non-significative	29
Snare Drum	**[tSa]**	**48**	[dum]	11
	[ta]	**34**	[pfit]	10
	[tu]	**18**	[pum]	7
	[ts]	**14**	[bum]	7
	[ti]	12	[tck]	5
	[S]	11	Non-significative	14
Bass Drum + Snare Drum Mixture	**[ta]**	**20**	[ts]	7
	[tSa]	17	Non-significative	12

Then, we designed a set of experiments to determine, from the most frequent onomatopoeia above chosen, which are the best to discriminate between snare and bass drum sounds by means of phonetic similarity.

Dataset The dataset chosen is composed of about 268 sounds (134 snare drums and 134 bass drums). The sounds were isolated strokes extracted from the *Freesound.org* Collection.

Language Model Since four sub-experiments have been performed, four different vocabularies have been designed. These are shown in Table 5.3. The grammar was the same for all of them: concatenation between vocabulary was not allowed. The vocabulary allowed are different combinations of the onomatopoeia selected after discussing the experiment conducted in [14]. Hence, for Exp. 1, given an input sound, the system is allowed to recognize [pum], [bum], [ta] or [tSa].

Table 5.3: Vocabulary Model for each of the four sub-experiments.

	Vocabulary Model
Exp. 1	bass: [pum]/[bum] snare: [ta]/[tSa]
Exp. 2	bass: [pum]/[bum]/[**tu**] snare: [ta]/[tSa]/[ts]
Exp. 3	bass: [pum]/[bum] snare: [ta]/[tSa]/[ts]/[**tu**]
Exp. 4	bass: [pum]/[bum]/[**tu**] snare: [ta]/[tSa]/[ts]/[**tu**]

Evaluation Each of the 134 bass and snare drum sound were input to the system. For each of these sounds the system determined the most likelihood onomatopoeia allowed in the vocabulary. For instance, if the output for a certain bass drum sound in Exp. 1 is [pum], this sound would count as 'O.K' (since [pum] is an onomatopoeia allowed for bass drums), but if the output would have been [ta], this sound would have been count as 'K.O.' (because [ta] is considered correct for snare drums but not for bass drums, as it has been decided at the begining of this subsection). Hence, by simply averaging among all the 'O.K' and 'K.O' for all the 134 bass drum sounds, we obtain a partial result of correctness for Exp. 1. Doing the same for the snare drum sounds, we finally obtain the total value of correctness of our system determining bass or snare sounds by phonetic similarity.

Results Table 5.4 shows the results for all four sub-experiments. Experiment 3 has the best results for the vocabulary model of 'SNARE' (82.83%), while for 'BASS' is Experiment 2 (79.85%). This can be explained by the presence of the onomatopoeia [tu] in both vocabularies. Hence, [tu] might be good descripting both snare and bass drums, but it also means that it would not be useful when the objective is to discriminate between both sound instruments. Finally, Experiment 4 performs the best results since both vocabularies have also their best combinations (81.34% for 'O.K' and 18.64% for 'K.O'). That is for bass drums the onomatopoeia [pum], [bum] and [tu]; for snare drums the onomatopoeia [ta], [tSa], [ts] and [tu] as well.

Table 5.4: Results for all four sub-experiments. Results are in percentages (%). Best result is obtained in Exp. 4, where the system is able to discriminate between bass and snare drum sounds the 81.34% of the times by means of phonetic similarity.

	partial BASS	**partial SNARE**	**TOTAL**
Exp. 1	O.K: 76.86 K.O.: 23.13	O.K: 58.2 K.O: 41.79	O.K: 67.5 K.O: 32.46
Exp. 2	O.K: **79.85** K.O.: 20.15	O.K: 46.27 K.O: 53.73	O.K: 63.06 K.O: 36.94
Exp. 3	O.K: 62.68 K.O.: 37.31	O.K: **82.83** K.O: 17.16	O.K: 72.76 K.O: 27.24
Exp. 4	O.K: **79.85** K.O.: **20.15**	O.K: **82.83** K.O: **17.16**	O.K: **81.34** K.O: **18.64**

5.3.3 Third Experiment

Initially given a certain onomatopoeia, the third experiment aims at finding the most likelihood sound to it within an heterogeneous sound dataset.

Dataset An heterogeneous dataset of about 190 sounds were randomly selected from the *Freesound.org* Collection. Different kind of sounds of different durations have been selected, such as animal sounds, field-recordings, instruments sounds, music loops, etc.

Language Model Depending on the given onomatopoeia, the vocabulary model is different. For instance, if the objective is to retrieve the most similar sound to a bass drum, so then, the vocabulary model would be composed of [pum], [bum] and [tu], (according to the results obtained in 5.3.2). Since in these experiments we are recognizing isolated words, loops are not allowed between words of the same vocabulary model.

Evaluation Given an initial onomatopoeia, such as [pum], all the sounds in the dataset are input to the proposed ASR system and their cost values are then compared. The sound with the lowest cost value might be the most similar one to the initial [pum].

Results First column in Table 5.5 shows the descending ranking of the most similar sound to a bass drum, while second column shows the output of the system. Hence, $34224 - HardPCM - Chip101$ would be the most similar sound to a bass drum, with output [tu]. In fact, it is perceptually very similar to a bass drum. The second most similar sound is also a bass drum, as the title indicates with 'bd'.

Table 5.5: This Table shows the results for the third experiment when looking for the most similar sound to a bass drum, presented as descending ranking. First column shows the sound title (under *Freesound.org* Collection nomenclature). Second column shows the system output.

Sound Title	**System Output**
34224-HardPCM-Chip101	[tu]
34442-anbo-bd-a-normal	[tu]
34208-acclivity-TongueClick1	[tu]
34249-pauliep83-agh	[tu]
34532-digifishmusic-Scrape	[pum]
34591-dobroide-20070506.rubber.duckies	[bum]
34168-Glaneur-de-sons-electric-wire	[tu]
34983-jonezy1476-Bell-2	[bum]
34330-marvman-hand-clap-comp	[tu]
34242-HardPCM-Chip127	[tu]

5.3.4 Fourth Experiment

Once validated the ability of the proposed system for finding the most similar sound in a dataset to an input onomatopoeia, the following was adding temporal considerations. Hence, in this fourth experiment loops among vocabulary are allowed. So then, phonetic evolution of sounds over time will be evaluated by observing the output phoneme sequences.

Two different sub-experiments have been performed. In he first one the vocabulary units are onomatopoeia such as [pum], and in the second one are single phonemes, adding thus a certain level of difficulty.

Dataset Some sounds of different databases have been evaluated. To present here we have selected some of the ENST-drum database [16].

Language Model Sub-Exp. 1 uses as vocabulary the onomatopoeia selected in the second experiment 5.3.2: [pum], [bum], [ta], [tSa], [ts] and [tu]. In Sub-Exp. 2, 23 single phonemes are the vocabulary units, as in experiment 5.3.1. Both grammars allow concatenation among respective vocabularies.

Evaluation The evaluation consisted on observing if the output sequence follows similar temporal pattern of each input sound. For instance, if an input sound consists on five strokes, observe if the system recognizes that by means of the onomatopoeia allowed in the vocabulary model.

Results Table 5.6 shows the results for Sub-Exp. 1. The system is able to recognize separated strokes (each [pum] or [ta bum] is separated by silence, '\|') in most of the cases. Only six examples have been selected to show in the following table.

Table 5.6: This Table shows the results for Sub-Exp. 1. First column indicates the sound title according the ENST nomenclature. Second column the number of strokes in each sound. Finally, third column shows the system output according to the vocabulary model. Symbol '\|' means 'silence' between two recognized onomatopoeia.

Sound Title	#	**Sytem Output**
015-hits-snare-drum-no-snare-brushes-x3	3	pum|pum|pum
035-hits-snare-drum-rods-x5	5	pum|pum|pum|tu
005-hits-bass-drum-pedal-x5	5	pum|pum|pum|pum|pum
006-hits-bass-drum-no-snare-pedal-x5	5	ts|pum|pum|pum|pum|pum
001-hits-snare-drum-sticks-x5	5	pum|ta bum|pum|ta tcha|pum
28-hits-snare-drum-no-snare-mallets-x7	7	pum|ta bum|pum|pum|pum|pum|pum

Results in Sub-Exp. 2 are presented in Table 5.7. Note that the difficulty has increased here towards Sub-Exp. 1, since the vocabulary units are 23 different single phonemes, instead of [pum], [bum], etc. As it can be observed the following table, the system is able to recognize phonetic patterns that suggest onomatopoeic representations, following somehow the temporal evolution of the sounds (compair the number of strokes and the number of different onomatopoeia).

Table 5.7: This Table shows the results for Sub-Exp. 2. First column indicates the sound title according the ENST nomenclature. Second column the number of strokes in each sound. Finally, third column shows the system output according to the vocabulary model. Symbol ’|’ means ’silence’ between two recognized strokes.

<table>
<tr><th>Sound Title</th><th>#</th><th>System Output</th></tr>
<tr><td>001-hits-snare-drum-sticks-x5</td><td>5</td><td>a|am|pam|pam|tam|pmam</td></tr>
<tr><td>020-hits-snare-drum-brushes-x5</td><td>5</td><td>um|bum|mum|u|m</td></tr>
<tr><td>022-hits-snare-drum-no-snare-mallets-x5</td><td>5</td><td>mu|mu|mu|pmu|pmu</td></tr>
<tr><td>016-hits-snare-drum-brushes-x7</td><td>7</td><td>pam|pt|pa|pam|buam|tmam|puam</td></tr>
<tr><td>028-hits-snare-drum-no-snare-mallets-x7</td><td>7</td><td>pmumu|am|m|mam|pmum|pbu|pmam</td></tr>
</table>

However, both approaches are still far from being used as classifiers of bass drum sounds against snare drum sounds (according to what it was obtained in 5.3.2), since results show that phonemes corresponding to both classes are used indiscriminately.

Chapter 6

Conclusions and Future Work

New ways of producing information, knowledge, and culture through social, rather than proprietary relations, are probably the main causes for the proliferation of online or social communities, aiming to collaboratively create multimedia databases. The context in which they are growing and many of their important aspects has been presented and discussed in this work.

The thesis focuses on some retrieval issues of sound effects in collaborative databases. In such repositories, high volumes of content are generated and annotated in an heterogeneous way by prosumers, usually users of online communities created around these databases.

Since it is impossible to automatically capture the way that humans perceive sounds, due to its subjective nature, [9], and as computers are still not sufficiently effective interpreting human descriptions or annotations of multimedia content, the effective management of online repositories is crucial for a correct retrieval and scalability.

This is basically the first attempt of the thesis: study the strengths and weakness of collaborative tagging and propose methods that can enhance the retrieval of audio content, specifically content which is scarcely annotated. A paper [34] concerning this was accepted in the Seventh International Conference on Machine Learning and Applications.

Alternative interfaces that allow different criteria when browsing and retrieving such large databases is the second focus of the thesis. Usually, large multimedia databases have different levels of descriptions. Hence, different users would then require different types of search, descriptions, etc of the multimedia content. Specifically, we propose a system that attempts to automatically extract timbre information at the phoneme level, by means of phonetic similarity. The starting hypothesis is that onomatopoeia or onomatopoeic representations are common ways to describe sounds that retrieval systems could take advantage. For instance, the keyword-based search can be thus complemented, allowing at the same time a more natural retrieval.

The following sections summarize the main conclusions of the thesis and suggest future work in order to improve both approaches.

6.1 Conclusions of the Freesound Experiments

The first part of the present thesis presents an analysis of the *Freesound.org* collaborative database, where the users share and browse sounds by means of tags, and content–based audio similarity search. We studied, mainly quantitatively, how users annotate the sounds in the database, and detected some well–known problems in collaborative tagging, such as polysemy, synonymy, and the scarcity of the existing annotations.

Regarding the experiments, we selected a subset of the sounds that are rarely tagged, and proposed a content–based audio similarity to automatically extend these annotations. Since the sounds in the test set contained only one or two rare tags, neither precision nor recall were applicable, so we used human assessment to evaluate the results. The reported results show that 77% of the test collection were enhanced using the recommended tags, with a high agreement among the subjects.

As future work, we are planning to extend the experiments using more sounds. In this case, automatic evaluation is needed. A possible solution is to select sounds belonging to similar sound categories (e.g all the percussive sounds scarcely annotated), and follow the same procedure of finding similar sounds from the *Freesound.org* database. So, the recommended tags should also belong to the same sound category. We are also working on a hybrid approach that combines textual similarity and content–based similarity to improve the recommendations.

6.2 Conclusions of the Phonetic Similarity Experiments

Sound search by phonetic similarity is the focus of the second part of the thesis. The motivation is the need of different search criteria when browsing large audio collections, since they may have many different levels of descriptions. By means of onomatopoeic representation of sounds, keyword-based search may be complemented, allowing at the same time a more natural retrieval.

The experiments and results presented in Section 5.3 are very promising although further research may be conducted in order to extend them to more sound categories, since percussive sounds have been the experiments focus, specifically, snare and bass drum sounds. Even they are very preliminary results towards a system that automatically captures phonetic similarity of non-speech signals, some conclusions can be extracted.

Humans use onomatopoeia, understood here as phoneme sequences, as a natural way to describe sounds. While semantic labels (such as *guitar*, *car* or *horse*) exploit high-level descriptions between word categories and the sounds they refer, onomatopoeic representations are related directly to acoustic sound properties. Both approaches can take advantage of each other. For instance, flexible interfaces could thus be designed, allowing different criteria for the sound retrieval.

Some other studies use manual approaches to annotate sounds with onomatopoeia, such as [52]. Often, manual categorization is affected by human subjectivity and thus such annotation can be error-prone and tedious, specially when databases are large. Hence, automatic techniques able to capture information at the phoneme-level are challenging and promising as well.

A very similar approach to ours is the one conducted in [24], where the authors first divided non environmental sounds into waveform chunks, under the theory that each peak in the signal power envelope corresponds to a onomatopoeic syllable. Then, each segment is analyzed by a HMM system, as ours, but this one previously trained with environmental sounds and their associated transcriptions (while ours is trained with an acoustic model of an open source ASR system). Similar low-level features are then extracted from each chunk and frame by frame. Finally, onomatopoeic representations are also constructed from each segment according to certain requirements of the Japanese language.

In contrast to other studies, we aim at designing a system which is language independant. This explains the reason of taking phonemes as system units (i.e. in the vocabulary model of the ASR system), instead of words. Hence, we explore phoneme concatenation capabilities (i.e allowing loops in the gramar of the ASR system) when describing timbre. Onomatopoeia expressions may vary due to cultural aspects, making thus hard their automatic recognition by systems which are just content-based. Then, another stage should be further included in order to map this acoustic-based information with common cultural understanding of familiar acoustic properties.

In addition to improvements that can be done into the system, audio segmentation could be applied. Examples that could benefit from this improvement would be those repetitive sounds where a single syllable is repeated over the whole excerpt. Then, the output of the system could be something like 'eight tik-tak'. Segmentation techniques could also help to refine the onomatopoeic representations from better detected events.

Finally, results suggest that the approach can be seen as a first stage in an system aiming at validating the use of Automatic Speech Recognition systems (ASR) for the automatic extraction of phoneme sequences in non-speech sounds. Specifically, some experiments show certain ability of the system to automatically extract timbre information by phonetic similarity in some sounds analyzed, this being a very promising result towards bridging the gap between acoustic and semantic sound descriptions.

List of Figures

List of Tables

Bibliography

[1] J. Andersen. *The Semantic Web Tutorial.* XML 2001, Finland.

[2] J.-J. Aucouturier and F. Pachet. A scale-free distribution of false positives for a large class of audio similarity measures. *Pattern Recognition*, 41(1):272–284, 2008.

[3] R. Baeza-Yates and J. A. Pino. A first step to formally evaluate collaborative work. *Proceedings of the international ACM SIGGROUP conference*, 35(12):56 – 60, 1997.

[4] R. Bechler. *Unbounded Freedom. A guide to Creative Commons thinking for cultural organisations.* Counterpoint. The cultural relations think-tank of the British Council., 2006.

[5] T. Berners-Lee. *Weaving the Web.* Harper, 1999.

[6] P. Cano. *Content-Based Audio Search from Fingerprinting to Semantic Audio Retrieval.* PhD thesis, Universitat Pompeu Fabra, 2007.

[7] S. M. Capito Silva, L. M. T. Jesus, and M. A. L. Alves. Acoustics of speech and environmental sounds. *Proceedings of ExLing-2006*, pages 221–224, 2006.

[8] C. Cattuto, D. Benz, A. Hotho, and G. Stumme. Semantic analysis of tag similarity measures in collaborative tagging systems. In *Proceedings of the 3rd Workshop on Ontology Learning and Population*, Patras, Greece, July 2008.

[9] B. E. and F. M. Audio information browsing with the sonic browser. In *Proceedings of IEEE Int.Conf.on Coordinated and Multiple Views In Exploratory Visualization (CMV'03)*, page 26, 2003.

[10] D. Eck, P. Lamere, T. Bertin-Mahieux, and S. Green. Automatic generation of social tags for music recommendation. In *Advances in Neural*

Information Processing Systems 20, pages 385–392. MIT Press, Cambridge, MA, 2008.

[11] T. Erickson and W. A. Kellogg. Social translucence: An approach to designing systems that support social processes. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 7(1):59 – 83, 2000.

[12] N. Fujisawa, S. ichiro Iwamiya, and M. Takada. Auditory imagery associated with japanese onomatopoeic representation. *Journal of PHYSIOLOGICAL ANTHROPOLOGY and Applied Human Science*, 23(6):351–355, 2004.

[13] W. Gaver. What in the world do we hear? an ecological approach to auditory event perception. *Ecological Psychology*, 5(1):1–29, 1993.

[14] O. Gillet and G. Richard. Drum loops retrieval from spoken queries. *Journal of Intelligent Information Systems*, 24(2-3):159–177, March 2005.

[15] O. Gillet and G. Richard. Drum loops retrieval from spoken queries. *Journal of Intelligent Information Systems*, 24:159–177, 2005.

[16] O. Gillet and G. Richard. Enst-drums: an extensive audio-visual database for drum signals processing. *Proceedings of ISMIR'06*, 2006.

[17] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry. Using collaborative filtering to weave an information tapestry. *Commun. ACM*, 35(12):61–70, 1992.

[18] S. Golder and B. A. Huberman. The structure of collaborative tagging systems, 2005.

[19] T. R. Gruber. A translation approach to portable ontology specifications, knowledge acquisition. *Knowledge Acquisition*, 5(2):199 – 220, 1993.

[20] A. Hazan. Towards automatic transcription of expressive oral percussive performances. *Proceedings of the International Concerence on Intellingent User Interfaces*, 2005.

[21] P. Herrera, A. Yeterian, and F. Gouyon. Automatic classification of drum sounds: A comparison of feature selection methods and classification techniques. In *Proceedings of the Second International Conference on Music and Artificial Intelligence*, pages 69–80, London, UK, 2002.

[22] P. Herrera-Boyer, G. Peeters, and S. Dubnov. Automatic classification of musical instruments sounds. *Journal of New Music Research*, 31, 2003.

[23] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. Information retrieval in folksonomies: Search and ranking. In *Proceedings of the 3rd European Semantic Web Conference*, pages 411–426, Budva, Montenegro, June 2006. Springer.

[24] K. Ishihara, Y. Tsubota, and H. G. Okuno. Automatic transformation of environmental sounds into sound-imitationwords based on japanese syllable structure. *Proceedings of EUROSPEECH-2003*, pages 3185 – 3188, 2003.

[25] J. A. Jacko and A. Sears. *The Human-computer interaction handbook fundamentals, evolving technologies and emerging applications.* Lawrence Erlbaum Associates, Mahwah, N.J., 2003.

[26] R. S. John Davies and P. Warren. *Semantic Web Technologies: Trends and Research in Ontology-based Systems.* Wiley, 2006.

[27] D. Jurafsky and J. H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition.* Prentice Hall, second edition, February 2000.

[28] A. Kapur, M. Benning, and G. Tzanetakis. Query-by-beat-boxing: Music retrieval for the dj. In *ISMIR*, 2004.

[29] W. Labov. The boundaries of words and their meanings. In *New Ways of Analyzing Variation in English*, pages 340–373. Georgetown U. Press, 1974.

[30] A. Lee, T. Kawahara, and K. Shikano. Julius – an open source real-time large vocabulary recognition engine. *Proceedings European Conference on Speech Communication and Technology*, pages 1691 – 1694, 2001.

[31] M. Levy and M. Sandler. A semantic space for music derived from social tags. In *Proceedings of 8th International Conference on Music Information Retrieval*, Vienna, Austria, 2007.

[32] B. Logan. Mel frequency cepstral coefficients for music modeling. In *Proc. of the 1st Annual International Symposium on Music Information Retrieval (ISMIR)*, 2000.

[33] B. Manjunath, P. Salembier, and T. Sikora. *Introduction to MPEG 7: Multimedia Content Description Language.* Ed. Wiley, 2002.

[34] E. Martinez, O. Celma, M. Sordo, and X. Serra. Extending sound folksonomies using content-based audio similarity. 2008.

[35] T. Masayuki, T. Kazuhiko, and I. Shin-ichiro. Relationships between auditory impressions and onomatopoeic features for environmental sounds. *Acoustical Science and Technology*, 27(2):67–79, 2006.

[36] T. Masui. Music composition by onomatopoeia. *Proceedings IWEC 2002*, pages 297–304, 2002.

[37] G. W. Matkin and F. H. Foundation. Technical report on learning object repositories: Problems and promise. Technical report, 2002.

[38] P. Mika. *Social Networks and the Semantic Web (Semantic Web and Beyond).* Springer, 2007.

[39] G. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information, 1956.

[40] T. Nakano, J. Ogata, and M. Goto. A drum pattern retrieval method by voice percussion. *Proceedings of the 5th International Concerence on Music Information Retrieval (ISMIR 2004)*, 2004.

[41] S. P. *Trait des objects musicaux.* Editions du Seuil, Paris, 1966.

[42] J. Paolillo and E. Wright. Social network analysis on the semantic web: Techniques and challenges for visualizing foaf. In *In A. Geroimenko and C. Chen, (Eds.), Visualizing the Semantic Web.* Springer, Berlin, 2005.

[43] A. D. Patel. *Music, Language, and the Brain.* NY: Oxford University Press, 2008.

[44] A. D. Patel and J. R. Iversen. Acoustical and perceptual comparison of speech and drum sound in the north indian tabla tradition: An empirical study of sound symbolism. *Proceedings of the 5th International Congress of Phonetic Sciences*, pages 925–928, 2003.

[45] J. Preece and D. Maloney-Krichmar. Online communities: focusing on sociability and usability. *The human-computer interaction handbook: fundamentals, evolving technologies and emerging applications*, pages 596–620, 2003.

[46] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[47] P. Resnick and H. R. Varian. Recommender systems. *Commun. ACM*, 40(3):56–58, 1997.

[48] H. Rheingold. *A slice of life in my virtual community.* MIT Press, Cambridge, MA, USA, 1994.

[49] B. Sigurbjörnsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. In *Proceeding of the 17th international conference on World Wide Web*, pages 327–336, New York, NY, USA, 2008. ACM.

[50] M. Sordo, C. Laurier, and O. Celma. Annotating music collections: How content-based similarity helps to propagate labels. In *Proceedings of 8th International Conference on Music Information Retrieval*, Vienna, Austria, 2007.

[51] S. Sundaram and S. Narayanan. Vector-based representation and clustering of audio using onomatopoeia words. In *Proceedings of AAAI 2006 Fall Symposia*, Arlington, VA, 2006.

[52] S. Sundaram and S. Narayanan. Classification of sound clips by two schemes: Using onomatopoeia and semantic labels. *Multimedia and Expo, 2008 IEEE International Conference on*, pages 1341–1344, 23 2008-April 26 2008.

[53] K. Tanaka. Study of onomatopoeia expressing strange sounds (case if impulse sounds and beat sounds) (in japanese),. *Transactions of the Japan Society of Mechanical Engineers*, 61(592), 1995.

[54] D. Tapscott and A. D. Williams. *Wikinomics: How Mass Collaboration Changes Everything.* Portfolio Hardcover, December 2006.

[55] A. Tindale. Annotated bibliography-drum transcription models. 2005.

[56] A. Tindale, A. Kapur, G. Tzanetakis, and I. Fujinaga. Retrieval of percussion gestures using timbre classification techniques. *Proceedings of the International Conference on Music Information Retrieval*, pages 541–544, 2004.

[57] D. Torres, D. Turnbull, L. Barrington, and G. Lanckriet. Identifying words that are musically meaningful. In *Proceedings of 8th International Conference on Music Information Retrieval*, Vienna, Austria, 2007.

[58] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet. Towards musical query-by-semantic-description using the cal500 data set. In *Proceedings of 30th International SIGIR Conference*, pages 439–446, New York, NY, USA, 2007. ACM.

[59] F. B. Viegas, M. Wattenberg, and K. Dave. Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 575 – 582, 2004.

[60] S. Wake and T. Asahi. Sound retrieval with intuitive verbal expressions. In *Proceedings of the International Conference on Auditory Display ICAD-98*, Glasglow, Scotland, 1998.

[61] J. Walker. Feral hypertext: when hypertext literature escapes control. In *Proceedings of the 16th conference on Hypertext and hypermedia*, pages 46–53, New York, NY, USA, 2005. ACM.

[62] Wapedia. Onomatopoeia, http://wapedia.mobi/en/onomatopoeia, 2008.

[63] S. Whittaker, E. Isaacs, V. O'Day, A. Adler, D. Bobrow, J. Bollmeyer, B. Damer, P. Dorish, T. Erickson, M. Jones, J. Larson, J. Li, W. Lutters, I. Paniaras, G. Rein, D. Sanderson, J. Sokolov, K. Tollmar, and C. Wolf. Cscw '96 workshop: widening the net: the theory and practice of physical and network communities: Nov.16–17, 1996, cambridge, ma. *SIGGROUP Bull.*, 18(1):27–32, 1997.

[64] WiseGeek. What is onomatopoeia, http://www.wisegeek.com/what-is-onomatopoeia.htm, 2008.