Read parts of this publication for more background on the FTT

Fourier-Time-Transformation (FTT), Analysis of Sound and Auditory Perception

Albrecht Schneider, Robert Mores

Schneider, A., & Mores, R. (2013). Fourier-Time-Transformation (FTT), Analysis of sound and auditory perception. In *Sound-perception-performance* (pp. 299-329). Springer, Heidelberg.
DOI: 10.1007/978-3-319-00107-4_13

For appreciation of the physiologically adequate approach to time-frequency-representations of sound and for the parameters of the FTT read these selected sections in the publication:

First part of	*1. Introduction*
For the FTT	*4. 'Perceptually adequate' Analysis and the Fourier-Time-Transform (FTT)*
and in particular	*4.2 Wavelets and FTT*

Please find an example demonstration of the FTT's time-frequency-representation for the challenging sound example (organ pipes together with bells) in Fig. 11, as compared to FFT in Fig. 10.

This is a citation of the publication and not a reproduction.

1. Introduction

In the present paper, we will reexamine the so-called Fourier/time transformation (FTT) that has been proposed by Ernst Terhardt (1985, 1992, 1998) as a tool for analysis and representation of audio signals such as speech and music. The main reason for suggesting such an approach was that Terhardt (1985) saw a different interpretation of the Fourier transform (as is widely used for spectrum analysis), on the one hand, and a need to develop a transform suited to perform time/frequency analysis comparable to that of the mammalian auditory system, on the other. Hence the aim of the FTT is to provide a time-to-frequency transformation equivalent to parameters in auditory processing as well as a "natural" approach to signal analysis (cf. Terhardt 1985, 1998, 78-97). In order to assess the possibilities the FTT approach might offer in regard to signal analysis, some other methods relevant for musical acoustics and psychoacoustics such as the short-time Fourier transform (STFT), autoregressive spectral modeling (AR) and Wavelet transform (WT) are presented in a brief survey, and are illustrated by some examples. Different approaches to time/frequency analysis are also viewed as to their power with respect to the so-called uncertainty product $\Delta t\ \Delta f$.

Over the past decades, there has been a broad range of research directed at understanding the functional anatomy and physiology of the auditory system (for summaries of research, see Oertel, Fay & Popper 2002, Pickles 2008, Winer & Schreiner 2011). Since about 1980, computational models of the auditory system have been issued that were progressively taking neurophysiological data and results from behavioral studies into account (for an overview, see de Cheveigné 2005, Meddis et al. 2010). By including elements representing hair cell transduction and neural activity patterns in the auditory nerve (AN) as well as in some of the relays along the subsequent neural pathway, complexity of the models as well as realism in performance has been increased by far (see, e.g., Meddis & O'Mard 1997, 2006). While most

current models are based in the time domain, there are some operating in the frequency domain. Traditionally, analysis in the time domain has been concerned with signal periodicity detection and estimation of 'pitch' from the repetition frequency of the envelope (f0). Analysis in the frequency domain typically has been done with the spectrum comprising a fundamental frequency f_1 and higher harmonics n x f_1 in view. For both approaches that have been pursued in auditory research for more than 150 years now (see de Boer 1976, de Cheveigné 2005), there are reasons at hand referring to the structure of audio signals (that can be represented both in the time and in the frequency domain) as well as with the functional anatomy and physiology of the mammalian auditory system. Considering only the first stages of auditory processing, and allowing for a rather schematic view, there is (1) transfer of waves from the environment through the ear channel to the tympanon. Then there is (2) a mechanical transmission line from the tympanon by means of the ossicles to the oval window where the pattern of vibration is transferred into (3) the cochlear fluid system in which a travelling wave with a relatively steep maximum for individual frequencies corresponding to sine tones is observed. Hence it has been concluded that a complex harmonic wave is decomposed in the fluid channel such that several maxima representing single partials or groups thereof will be observed. The cochlear partition with (4) the basilar membrane (BM) and as well as structures combined with the BM are regarded as a filter bank of k channels capable to decompose a complex signal into partials or groups thereof. (5) Inner hair cells (IHC) effect mechanoelectrical transduction so that the output of each of the BM channels is coded into a train of neural spikes that are (6) represented in fibers of the AN. Modeling transmission of audio signals from the pinna to the stapes (a mechanical system with impedances and admittances) and within the fluid ducts of the cochlea (a hydromechanical system that incorporates nonlinearities; see Nobili & Mammano 1996) as well as the transduction mechanism on the IHC and AN level is quite complex since every element in the transmission chain as well as their interaction must be adequately covered, that is, as close as possible to empirical data from (mostly, animal) experiments and behavioral studies (cf. Meddis & Lopez Poveda 2010).

In regard to such a complex transmission line that may incorporate also relays of the auditory pathway such as the cochlear nucleus (CN) or models for processing at even higher levels (the superior olivary complex and the inferior colliculus), restricting an analysis to peripheral filtering processes as effected in the cochlea (as is done in this paper) may seem odd. The point, however, is that initial analysis on the BM and IHC level seems decisive since it can be shown that distinctive features of complex sounds such as salient or ambiguous pitch structure, harmonic or inharmonic spectrum (leading to percepts classified as consonant or dissonant), and also phenomena such as combination and difference tones are derived from peripheral processing (for examples, see Schneider & Frieler 2009). In the case of the peripheral processing lacking sufficient precision (consequent to, for example, inappropriate design of BM filters), feature extraction at this stage of processing and also on higher levels of the auditory pathway can be significantly hampered.

2. Uncertainty Relation and Time/Frequency Resolution

The uncertainty relation known from quantum mechanics states that a particle can be defined exactly either as to its impulse p or to its place x. Since exact definition of the impulse

precludes exact definition of the space (in regard to wavelength), a situation where both have to be taken into account leads to the product of place and impulse such that $\Delta x\ \Delta p \geq \hbar/2$ ($\hbar = h/2\pi$ with h = Planck's constant). This basic equation became known as the uncertainty relation and has been adapted, with necessary modifications, into various fields of science such as communication theory and acoustics (Gabor 1946).
According to Gabor (1946), for signals a limit for the product of time resolution and frequency resolution exists like

(1) $\Delta f\ \Delta t = 1/2$

This minimum is restricted to very few 'ideal cases' (see below) so that for real signals such as sound of a certain duration and bandwidth values above 0.5 will apply. In a general formulation, the uncertainty relation for acoustic phenomena such as impulses (cf. Meyer & Guicking 1974, 92ff.) can be given as

(2) $\Delta t\ \Delta f \geq 1$

As can be demonstrated by calculation, the lower limit of $\Delta t\ \Delta f = 1$ can be achieved for a Gaussian impulse while for almost every other pulse type $\Delta t\ \Delta f > 1$ applies.

Taking two extremes, a Dirac-δ (with a duration approaching zero and an impulse height approaching infinity) and a sine wave of an arbitrary frequency f_i lasting from $-\infty < t < \infty$, the impulse is defined exactly as to time t (ms), and the sine wave as to frequency f (Hz), in a two-dimensional time-frequency space. "Real-world" signals such as produced by musical instruments including the human voice are neither as short in duration as a Dirac-δ, nor infinite in duration as the undamped sine wave repeating itself at the same frequency. Of course, in regard to spectral bandwidth, the Dirac impulse and the sine tone of a given frequency also represent two extremes. In music as well as in other audio signals such as human speech or birdsong, the situation typically is that a number of complex sounds each comprising n harmonic or inharmonic partials occur at a certain time, and have disappeared due to damping forces after a duration of, in most cases, a few hundred milliseconds or perhaps several seconds. Hence we are dealing with sequences of complex sounds such as melodies, or with several such sequences played or sung more or less in parallel (in regard to tracks of fundamental frequencies) as well as more or less synchronous (as regards onsets of tones/notes) as in homophonic and polyphonic music.

In this respect, conventional western staff notation constitutes an acceptable approximation to a two-dimensional time/frequency representation with the ordinate y giving frequency on a log scale, and the abscissa x time on a linear scale (cf. Rossing 1982, 134-135). One can therefore substitute staff notation with semi-logarithmic graph paper to yield a similar (but more precise) notation for monophonic or polyphonic music (for an example of a Bach chorale with four voices, see Schneider 2001). It has to be noted, in this context, that western staff notation in regard to 'pitch' information represents the fundamental frequency f_1 (as is obvious from definitions such as standard pitch A_4 = 440 Hz or "middle c" [C_4] = 261.6 Hz in equal temperament). Whether the tone notated on staff as C_4 is a pure (sine) tone or a complex tone cannot be gained from Western staff notation, which does not include spectral information. However, it is implied from A_4 = 440 Hz that any complex tone played to render

this note audible should comprise a fundamental frequency f_1 at 440 Hz (though, at least in perception, a 'pitch' corresponding to 440 Hz could be realized also with an envelope repetition frequency f0 = 440 Hz while the fundamental of the spectrum is weak or even missing).

Of course, one could further substitute staff notation with a melogram or spectrogram (sonogram) as a two-dimensional representation of sound and music in a time/frequency space. We will do this with a musical example offered recently by Florian Messner (2011) who, together with another singer, recorded a phrase noted down in staff notation by Franchino Gafori (Franchinus Gaffurius, 1451-1521), in his *Practica musicae* (Milan 1496, Gafori (Lib. III, cap. 14: de falso contrapuncto) gave us this piece of two-part music then still in practice in the Lombardic in vigils and in the mass for the dead because he thought it defied all rules of counterpoint (…*ab omni modulationis ratione seiunctus est*). What in fact singers were performing was vocal music where two voices go in parallel with dissonant intervals (seconds, fourths) between them. Singing styles as well as instrumental music organized as a diaphonia with two voices forming narrow intervals were or even still are in use in the Balkans (notably in areas of Bosnia and Herzegowina, Croatia, Albania, Bulgaria). Since two notes sung in parallel at the interval of a minor or a major second will have fundamental frequencies so close as to fall into one 'critical band' (CB), they cannot be separated by the auditory filter bank, and thus a sensation of roughness from the interaction of fundamental frequencies as well as from other partials in their respective CBs will result. In Bulgarian diaphonic singing, one finds two (female) voices approaching each other as close as ca. 45 - 80 cents (cf. Schneider et al. 2009), that is, from about a quarter tone to a chromatic semitone.

For the Lombardic *contrapunctus falsus* as performed by two male singers, the spectrogram shown in figure 1 results.

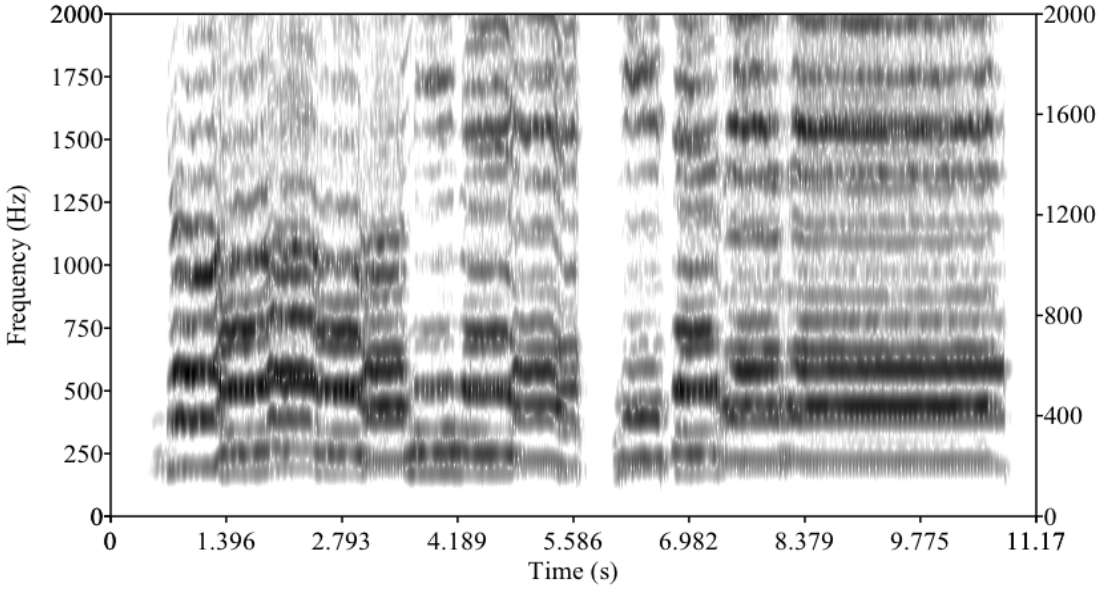


Figure 1: Lombardic diaphony, two male singers, spectrogram 0 – 2 kHz

Though the spectrogram has been calculated in the frequency domain with a rather high resolution as to time and frequency[1], the trajectories of the fundamental frequencies for the two voices will be difficult to recognize. Also kind of a melogram representing the pitches (calculated in the time domain with a special autocorrelation algorithm, Boersma 1993) will give only some rough idea as to the movement of the voices (see figure 2):

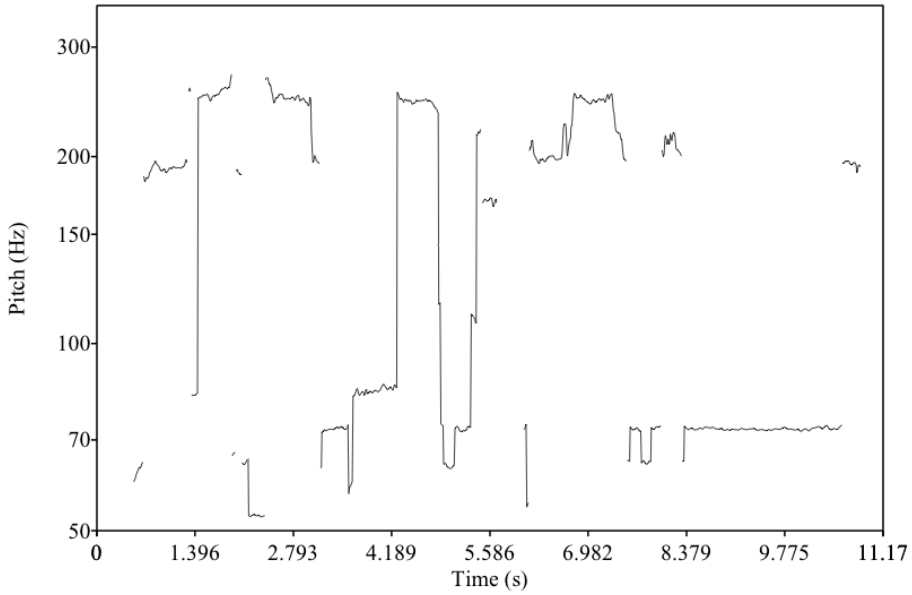


Figure 2: Pitch (f0) tracking for Lombardic diaphonia, autocorrelation method

It is possible to find the fundamental frequencies for the two male voices even for narrow intervals with a standard frequency analysis based on FFT, provided the window of analysis is long enough to ensure that relevant components can be separated.

Applying a Discrete Fourier Transform (DFT, cf. DeFatta, Lucas, Hodgkiss 1988, 238ff.) to a digital signal $x(n)$ with a period of T, the frequency resolution Δf depends on the sampling rate F_s and the transform length (often also called 'frame' or 'window') of size N. The discrete frequencies f_k for a spectrum $X(k)$ of the signal can be calculated as

(3) $f_k = k\ (F_s/N)$ where $k = 0,1,2,3,\ldots, N\text{-}1$ is the frequency index.

The frequency resolution hence depends on the ratio F_s/N and can also be expressed as

(4) $\Delta f = 1/T = F_s/N$

It is obvious from equation 3 that basic relations defined for analogue band pass filters hold likewise in the digital domain. For a narrow-band filter (cf. Küpfmüller 1968, 71f.), the response time τ is defined as

(5) $\tau = 2\pi/\Delta\omega = 1/\Delta f$ (for $\omega = 2\pi f$)

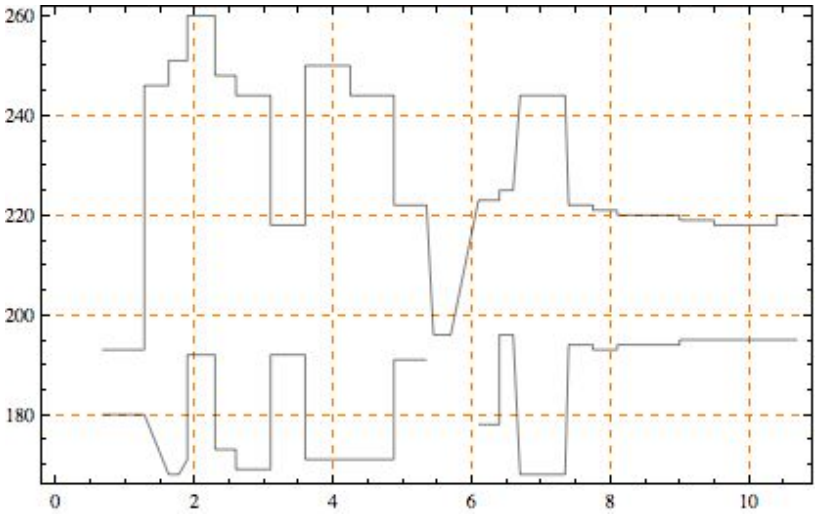
[1] Settings for the analysis performed with the Praat software (Boersma & Weenink 2011) were a time window of 30 ms with a Gaussian weighting, a time step of 2 ms from one frame to the next, an analysis bandwidth of 2 kHz and a frequency step of 2 Hz. The sound sample of 11.17 seconds was processed in 5253 (overlapping) frames.

Hence the response time and bandwidth of the filter are in reciprocal relation. For any frequency resolution Δf designed for the filter, a corresponding response time τ can be calculated; since τ in this respect defines Δt of the filter (taken as an ideal, non-dispersive band pass; cf. Meyer & Guicking 1974, 92ff., 346ff.), the product $\Delta f \, \Delta t \geq 1$ applies equivalent to Equ. (1)[2]. The uncertainty relation that, as a general principle, needs to be adopted for specific areas, underlies also digital sampling and frequency analysis (Equ. 2, 3) where a signal $x(n)$ of period T sampled at F_s can be determined in regard to its spectrum $X(k)$ the better the longer the transform size N is chosen. This, however means that good frequency resolution Δt can be achieved only at the cost of rather poor time resolution Δt.

With respect to our example, the Lombardic diaphonia, the sample rate of 44100 per second will require a window size or transform length of at least $2^{12} = 4096$ to ensure a frequency resolution $\Delta f \sim 10.77$ Hz. As can be easily checked, the exact value for Δf is 10.7666 Hz; Δt is determined by the transform of length $N = 4096$ samples = 92.8798 ms. If we leave out windowing and other effects, the product of time and frequency achieved in FFT-based analysis indeed would be unity[3]. For the analysis of the sound example, FFT windows of 2^{12}, 2^{13} and 2^{14} samples were employed together with a spectral peak estimation algorithm. Frequency readings were confined to full frequency values (e.g, 195 Hz, 222 Hz) averaged over the window of length N. The results of the time/frequency analysis have been tabled and then plotted as shown in figure 3. For reasons of readability, a linear frequency scale (ordinate) was chosen. The movements up and down (melodic contour) as well as musical intervals formed between the two voices by their fundamental frequencies over time are clearly visible. However, the relatively poor time resolution of the analysis is also quite obvious since the 'pitches' sung (represented by their respective fundamental frequencies f_1) are indicated according to the transform size that has been employed. For example, at F_s = 44100 samples, a window of 8192 samples means a time interval of 185.76 ms for which a spectrum is calculated that contains information as to the 'average pitch' that, in our example, was realized by two singers within this span of time. In reality, there can be marked shifts of fundamental frequency within one frame or window of length N. In fact, the intonation practiced by the two singers in recording this piece of music shows far more subtle fluctuations than shown in figures 2 and 3 as became obvious in a more detailed analysis carried out with high resolution tools (Wigner transform and FFT combined with LPC pitch tracking and very small hop ratios).

[2] A formal proof can be given on the basis of the Cauchy-Bunjakowski-Schwarz inequality (cf. Meyer & Guicking 1974, 95, 108; Papoulis 1962, 63).

[3] Applying no specific windowing function means a rectangular window is chosen for which the so-called Equivalent Noise Bandwidth (ENBW [Bins], see DeFatta, Lucas, Hodgkiss 1988, 262ff.) is 1.0.



Words: Do – o – o – o – o – mi – ne, mi – se – r e – r e

Figure 3: Lombardic diaphonia, 2 male voices, tracks of fundamental frequencies ./. time

What is evident from figure 3 is that the two singers didn't start in unison (what the notation provided by Gafurius would have demanded) but at an interval of about a semitone (193 : 180 Hz ~ 122 Cents). Also, one can see that at the end of the phrase (from 7.5" to 10.7" on the time scale) a long dissonant interval, namely a major second based on the notes G_3 and A_3 occurs. While singing their respective notes/tones forming the major second, the singers adjust their intonation several times (the interval size varies from an initial 233/234 cents to ca. 201 and even 193 cents towards the end). There are some more details one can study with the data condensed in figure 3 at hand. Figure 3 can be regarded as kind of a descriptive 'notation' derived ex post from an actual performance. This notation, by the way, could be transformed back into a symbolic notation (e.g., western staff notation).

If one would need to improve temporal resolution of the analysis, there are methods at hand in digital signal processing (DSP) which permit to achieve this goal without suffering adequate frequency resolution. One of the most basic and at the same time most efficient procedures is to overlap consecutive frames of analysis (what has been done to some degree also for the present analysis). In case overlap is almost complete and the so-called 'hop ratio' therefore very small, a sequence of signal spectra will result following one another at a short delay of n samples while the frequency resolution of each spectrum is determined by N. Such an analysis technique is well suited for transients where the rate of change in the signal per time often is significant. We will show an example for such an analysis below. The point of interest with respect to choosing a certain method of analysis of course is this: what is the degree of exactitude necessary in regard to (a) auditory perception and relevant psychoacoustic parameters? Further, which technique should be used if (b) the study of musical structure is an issue (e.g., when studying music not well documented yet)? In addition, signal analysis could also be pursued in regard to (c) acoustics of certain instruments where the aim often is to investigate processes of vibration, sound production and sound radiation. The precision needed under (c) is certainly much higher than that required for (a) or wanted for (b).

Taking figure 3 as an example, one may call the analysis plausible in regard to musical structure since the melodic contours of the two voices and the intervals formed between them can be followed with ease. What is less accessible to intuitive understanding in this plot, though, is the exact size of the intervals realized by the two voices. Of course, musicians and musicologists will have an idea as to the fundamental frequencies of notes in a diatonic scale (at least in regard to main intervals). However, a number of deviations in intonation that were documented in the signal analysis are difficult to read from the tracks in figure 3. In regard to auditory perception, the precision achieved in the plot in figure 3 probably is above that ordinary listeners might achieve by using their ears only for analysis (even trained musicians might find it difficult to separate the two voices which are quite close in register, and in the recording at hand do not differ much as to their respective timbre). In sum, one could argue that the analysis as shown in figure 3 is sufficient to illustrate a musical structure as was put to sound by two male singers, and it represents about the result trained listeners might obtain from an aural analysis of the musical phrase as recorded on CD.

In regard to time and frequency resolution as are most relevant for signal analysis, it should be noted at this point that the 'uncertainty relation' (or 'relation of indeterminacy') yields $\Delta f\,\Delta t \geq 1$ for linear systems such as analogue band filters[4]. For the auditory system, it has been shown in experiments based on biophysical cochlea models (cf. Mammano & Nobili 1993, Nobili & Mammano 1996) that time/frequency analysis of the cochlea for the range of speech signals above 200 Hz already for a passive model comes close to $\Delta f\,\Delta t \approx 0.55$ (Russo, Rožić & Stella 2011), that is, very close to the theoretical limit of 0.5 as defined by Heisenberg's 'uncertainty relation' or the equivalent formulation Gabor (1946) has given for time/frequency resolution as a relevant parameter for communication systems. The general concept Gabor advanced was that for every type of resonator a *characteristic rectangle* of about unit area can be defined in a time/frequency plane. For a sharp resonator such as a narrowband filter $\Delta f\,\Delta t \approx 1$ can be assumed. From mathematical considerations as well as from properties of some elementary signals (sine or cosine wave, Dirac-δ) Gabor (1946, 435) concluded that the signal for which $\Delta f\,\Delta t = 1/2$ *applies is the modulation product of a harmonic oscillation of any frequency with a pulse of the form of a probability function*. (For an 'ideal' bandpass filter he calculated the value 0.571). Gabor suggested that a time/frequency space (understood as an information diagram with the axes time and frequency) can be divided into rectangles which have sides defined by Δf and Δt, respectively. According to Gabor, each area $\Delta f\,\Delta t$ *represents one elementary quantum of information*; he therefore proposed to call such an area a *logon*.

Remarkably, Gabor (1946, Part 2) included hearing into his study, where he is making reference to several empirical studies on difference limens for pitch and time (as had been published by Shower and Biddulph in 1931, and by Bürck, Kotowski and Lichte in 1935; see below). Gabor argued that the ear (or, rather, the sense of hearing) disposes of a *threshold information area* in regard to frequency (pitch) and time, and of an *adjustable time constant*

[4] There are several definitions as to ‚linear'. In electronics, linear refers to circuits (like LRC filters) in which linear relations exist between physical magnitudes (induction, capacity, resistance, gain) and where all voltages and current are proportional to the electromotive force driving the system (cf. Küpfmüller 1968, 12f.). In signals and systems theory, linearity is defined by Bachmann (1992, 9) like this: superposition at the input has the same effect as superposition at the output.

at least between 20 and 250 ms. Thus he regards hearing a most relevant field where his concept of time/frequency areas or *logons* is of practical significance.

It is obvious that basic ideas as formulated by Gabor for signal and systems theory also underlie some other approaches, notably wavelet analysis (cf. Dutilleux et al. 1988, Mertins 1999, ch. 7, Evangelista 1997). In fact, it can be demonstrated that, in regard to fundamental mathematical concepts, formal equivalence exists for the Wigner transforms, Gabor coefficients, and Weyl-Heisenberg wavelets (see Dellomo & Jacyna 1991). Gabor's concept and related concepts by Eugene Wigner and J. Ville have led to a systematic treatment of linear and non-linear time/frequency analysis of signals (see Cohen 1995, Flandrin 1999, Mertins 1999). Application of the Wigner transform (WiT) to acoustical signals is possible with some modification of the original formulation (cf. Yen 1987) and can yield high-resolution time/frequency representations. For a complex-valued signal $\boldsymbol{s}(t)$, WiT can be calculated according to

$$(6)\ W(t,\omega) = \int_{-\infty}^{\infty} e^{-j\omega\tau} s(t+\frac{\tau}{2}) * (t-\frac{\tau}{2}) d\tau\ ,$$

where * denotes the complex conjugate. For practical applications in DSP, the integral comes down to a summation, and a window function is applied since the WiT is a bilinear transform that produces cross terms between spectral energy peaks resulting from a real-valued signal. The cross spectrum appears in the time and in the frequency representation and contains sum and difference of the original spectral components. The window function helps to cancel out cross terms. Also, a good comprise solution suited to suppress spurious spectral components is a combination of FFT and WiT for which parameters can be set so as to cancel out most of the unwanted cross terms while improved resolution (as compared to FFT alone) is maintained.

As an example …

3. Time/frequency analysis: some applications and examples

There are …

4. 'Perceptually adequate' Analysis and the Fourier-Time-Transform (FTT)

In the following, some fundamentals of psychoacoustics will be considered and compared to parameters found in DSP-based analysis and auditory modeling. The latter aims at a realistic 'emulation' of the auditory system in regard to basic functions and actual performance (cf. Meddis et al. 2010). Signal-analysis tools such as WT and FTT are less complex than full-grown auditory models (e.g. Meddis & Lopez-Poveda 2006), however, they can be viewed as representing the initial stage of BM filtering and thus are important as auditory 'preprocessors' (cf. Solbach, Wöhrmann, Kliewer 1998, Terhardt 1998) that generate output used further in pitch and loudness perception as well as in auditory scene analysis. It should be underpinned that effective neural processing of complex sound naturally depends on the quality of (peripheral) BM filtering; the faster and the more precise this stage operates, the better neural processing along the auditory pathway can be achieved.

4.1 Frequency and time resolution; discrimination and recognition tasks

The Fourier integral (see Bracewell 1978, ch. 2, Meyer & Guicking 1974, 70ff.) which is fundamental to Fourier analysis can be viewed as presenting a time function $x(t)$ in terms of frequency (or, rather, angular frequency ω). The Fourier integral considers frequency in an infinite interval ($-\infty \leq T \leq \infty$) and thus, as Gabor (1946, 431) has put it, *sub specie aeternitatis*. In musical signal analysis, however, one has to work with sounds that change over time, and often abruptly so. The answer to this situation was to consider applicability of Fourier theory to signals of definite length as well as to signals that lack clear periodicity and which are inharmonic in spectral composition. For practical reasons, techniques such as STFT (see Mertins 1996, ch. 4; 1999, ch. 7) were developed. The basic concept for STFT is to multiply a sound signal $x(t)$ by an analysis window $g(t)$ and then compute the Fourier transform. For the analysis of a time signal, typically windows of length $N = 2^n$, $n = 8, 9,\ldots, k$ are chosen. If the signal to be analyzed is longer than N, the signal is processed frame by frame (with an overlap of 50% or more to ensure continuity). Hence the window "slides" along the time axis by an amount defined by a shift parameter τ. The result thus obtained can be displayed in 2D or in (quasi) 3D-images such as figure 5 above. Though the STFT is regarded a good analysis tool that has been widely applied in acoustics and in particular in musical acoustics, it has a certain disadvantage in that conventional Fourier-transform algorithms operate on fixed values for N, which defines both Δf and Δt in a two-dimensional time/frequency plane (with f [Hz] as ordinate and t [ms] as the abscissa). Hence, time and frequency resolution are constant over the total bandwidth of analysis. In terms of Gabor's *logons* (see above), a uniform rectangle as "analysis box" results for low as well as for high frequency bands. An analysis window of constant length $N = 2^n$ samples applied to the full bandwidth of human auditory perception (ca. 25 Hz – 16 kHz) seems unfortunate because our auditory system apparently needs a certain number of signal periods rather than a fixed time interval for pitch analysis (see below). Since the period duration T (ms) varies with frequency, the analysis window (either expressed in ms or in the number of samples) should be longer for low frequencies as compared to middle and high frequency bands.

In regard to temporal resolution relevant to hearing, a range of 'time constants' basic to temporal integration has been issued. It has been critically remarked that *"time constants" estimated from different experimental tasks range over three order of magnitude, from 250µs to 200.000µs* (Eddins & Green 1995, 207). In fact, there are different time constants relevant for different perceptual tasks as well as in regard to triggering motor responses, etc. In view of acuity achieved in discrimination tasks, minimum integration time in hearing appears to be 2-5 ms, depending to some extent on types of stimuli and conditions (see, e.g. Bilsen & Kievits 1989 who used so-called white flutter pulses). The data, which have been obtained in gap detection as well as in other experiments, are uneven (cf. Moore 2008, ch. 5). Among relevant factors, time-intensity trades have to be taken into account (temporal integration depends on intensity or sound level; see Eddins & Green 1995). If minimum integration time of ca. 2-5 ms is interpreted in terms of response time of the auditory filter (as has been done), it appears that the response time perhaps plays a small role at low frequencies ($100 < f_{gr} < 500$ Hz) but not for frequencies above 1 kHz.

Other ‘time constants’ refer to noticeable asynchronies in the onset of the same tone played by two instruments (typical values seem to be $10 < t < 20$ ms), to “smearing” of several discrete echoes that occur in a room within a certain time span ($t < 50$ ms) into a sensation of quasi-continuous reverberation, and to temporal integration of energy in the sensation of loudness (most experimental data suggest an interval of $100 < t < 200$ ms). In regard to such ‘time constants’, one of course has to distinguish between discrimination and identification tasks, not to forget temporal organization of sound objects on a higher level such as grouping and chunking in music cognition (see Snyder 2000). Discrimination for example in 2fc-experiments simply calls for responding if a certain ‘event’ did happen or not irrespective of what the informational ‘content’ of such an event may be. A very short pulse or noise burst will be sensed as a ‘knack’ but is not accessible for detailed auditory analysis. Even decisions subjects have to make whether a stimulus presented in a pair of sine tones is ‘higher’ or ‘shorter’ than the other (a design typical of experiments directed to difference limens for Δt and Δf relative to frequency bands) might just require a modicum of information on the side of the subject as to the nature of the stimuli. In contrast, identification of a stimulus in regard to one or several properties needs considerably more time since sound input that has been transformed into neural spike trains must be processed along several stages of the auditory pathway before, for example, a certain ‘pitch’ can be assigned to a stimulus. If one accepts periodicity detection and temporal processing for pitch as the predominant principle (notwithstanding significant evidence for rate-place representations and tonotopicity), the periods of time signals that might occur in musical sound are roughly from 33 ms (30 Hz) to 0.067 ms (15 kHz). Therefore, as maximum lag of 33 ms has been implemented in an ACF model suited to account for very low frequencies down to 30 Hz (Pressnitzer, Patterson, Krumbholz 2001). In addition, time needed for arbitrary pitch estimates has been suggested as being 66 ms, with possibly less time down to about 40 ms or even 20 ms needed for such signals where subjects have a certain knowledge as to their likely pitch range in beforehand (cf. de Cheveigné 2005, 205). If 66 ms is a correct ‘time constant’, for most of musical relevant frequencies it would cover several or even many periods. In some early experiments, the time needed for developing a clear sensation of pitch for a sine tone varied from about 60-100 ms for very low frequencies (50 Hz) and ca. 30 ms for 300 Hz to about 15 ms for a frequency range of ca. 0.5 to ca. 5 kHz (Bürck, Kotowski, Lichte 1935). From the empirical data as well as from considerations concerning the physics of the signal (that was switched on and off in an electronic circuit) and conditions of measurement, Bürck and colleagues calculated curves of tone recognition times as a function of frequency where about 80-100 ms would be required for a sine tone of 100 Hz but only ca. 5-10 ms for a sine tone in the range 1-5 kHz. Taking these approximate figures, one may hypothesize that pitch estimates for sine tones require about 5-8 periods of the time signal. The estimate figures mentioned above (to which several more from various experiments can be added) can be taken as tentative time constants in computational models of auditory perception.

In regard to frequency discrimination in hearing, for frequencies of two pure (sine or cosine) tones presented one after another, and with constant sound pressure level (SPL), the difference limen (DL) or just noticeable difference (jnd) has been estimated to be of the order of 1/30 of the Critical Bandwidth (CB). The concept of CB (see Moore 1995, Zwicker & Fastl 1999, ch. 6) refers to BM excitation and filtering. From empirical data, a cochlear tonotopic

frequency map has been proposed (cf. Greenberg 1990) where one CB corresponds to ca. 0.89 mm of BM. Hence, 1/30 of this unit would have to be considered as the jnd in regard to place theories of pitch and BM excitation patterns. However, one has to see that hearing is a dynamic process based on feedback regulation and fast adaptation to stimulus conditions (otherwise, extremely sharp frequency discrimination as observed in trained musicians and very short recognition times for pitch and timbre of complex sounds would not be possible). Therefore, it seems only natural to see that center frequencies, bandwidths and shape of auditory filters (AF) vary with BM excitation level and bandwidth of input signals. Further, it is obvious that CB models such as have been proposed for loudness summation and place theories of pitch should be taken as a basic concept that must be validated with empirical data since a number of assumptions pertaining to CB models do not hold in a strict sense (cf. Moore 1995). Empirical data on CBs indicate that the Bark scale comprising 24 or 25 (in theory: non-overlapping) filter bands is not quite appropriate in particular for low frequencies (f_c < 500 Hz) since the bandwidth of the AF increases significantly with decreasing frequency. This effect is most prominent for f_c < 200 Hz (cf. Jurado & Moore 2010, Schneider & Tsatsishvili 2011). Compared to the Bark scale (cf. Zwicker & Fastl 1999), the so-called ERB scale (ERB = Equivalent Rectangular Bandwidth) comprising about 40 filter bands fits better to perceptual data though it does not fully account for pronounced increase of bandwidth at low frequencies. Each ERB is calculated by taking 4 f_c/p, where f_c is the center frequency and p is a filter parameter that determines the passband and the slope of the filter. In regard to modeling, the "effective bandwidth" for each AF along the BM depends on place and center frequency (that apparently is not fixed yet variable within a certain range), on sound level as well as on spectral energy distribution and spectral flux within audio signals. Very roughly, one can approximate CBs by 1/3 octave band pass filters. In reality, the "effective bandwidth" of AFs seems to vary from about one octave at very low frequencies to close to 250 cent around 1-3 kHz.

4. 2 Wavelets and FTT

Wavelet analysis is one of several methods that have been developed to account for Gabor's *logon* concept and to provide equally good time and frequency resolution over the bandwidth of auditory perception. Wavelet analysis basically can be viewed as a Fourier approach where the window of analysis $g(t)$ is shifted in frequency by Ω_0, that is, multiplied in the time domain by $e^{i\Omega_0 t}$. Similar to STFT, a sliding process along the time axis is part of the analysis with an increment of τ. Wavelet analysis (cf. Dutilleux, Grossmann, Kronland-Martinet 1989) further includes a part equivalent to the 'window' $g(t)$, namely the analyzing wavelet $h(t) = e^{i\Omega_0 t}\, g(t)$ that is dilated in frequency by a parameter a so that

$$h^{(a,\tau)}(t) = \frac{1}{\sqrt{a}} h\left(\frac{t-\tau}{a}\right) \tag{9}$$

The wavelet transform (WT) of a continuous time signal $s(t)$ then is

$$W_h(\tau, a) = \frac{1}{\sqrt{a}} \int h\left(\frac{t-\tau}{a}\right) s(t)\, dt \tag{10}$$

The wavelet transform is computed by convolving the signal with a time-reversed and scaled wavelet (see Evangelista 1997). In regard to sound analysis, WT can be considered as a kind of band pass filter where the center frequency and the bandwidth of the filter can be varied by different values for the parameter *a* (cf. Mertins 1999, ch. 9). In this respect, WT effectively computes a *constant-Q* filter analysis as has been employed in the gammatone filter analysis shown above (figure 9) where WT was performed for a frequency band of 0 – 1.6 kHz divided into four octaves each of which was subdivided into four bands of 250 cents to approximate the bandwidth of the auditory filter (AF) with respect to CB concepts.

A concept similar to STFT as well as to WT in certain respects is the Fourier-Time-Transform (FTT) as proposed by Terhardt (1985). In an article in which he considered properties of several different Fourier transforms, Terhardt argued that Fourier transforms are not restricted to periodic signals, and that the actual analysis window must not be identical with a period (or several periods) of a time signal $p(x)$ to yield valid spectral representations (a criterion to check validity of course is whether or not restoration of the time signal from the spectral data by an inverse transform can be achieved). Without going into details (many of which relate to linear systems theory rather than to "plain" spectral analysis), the argument put forward by Terhardt is that, for causal systems and signals, analysis of a physical signal such as sampled sound can be confined to time intervals from $t = 0$ to t so that the FTT for one-sided signals is given by

$$(11) \qquad P(w,t) = \int_0^t p(x)e^{-wx}dx \,;\; t > 0 \text{ and } w = \mathrm{j}\, 2\pi f = \mathrm{j}\omega$$

The spectrum $P(w, t)$ for every instant t represents the time signal within a time interval that is defined as $-\infty < x \leq t$. Also, $p(x) = 0$ for $x < 0$. For practical applications, signal values that are far in the past are of little relevance as to the current state of a system or signal[5]; therefore, the signal is multiplied by an exponential weighting function $\exp(-a(t - x))$ where $a \geq 0$ is a damping factor that can have values of 0 to 1. Consequently, with the exponential weighting included, Equ. 8 becomes

$$(12) \qquad P(w,t) = \int_0^t p(x)e^{-a(t-x)}e^{-wx}dx \;;\; t > 0$$

FTT applied to one-sided signals yields two parts, one steady-state and one transient (cf. Terhardt 1985, equations 32 and 33)[6]; the transient part vanishes with ongoing time; also, amplitude density distribution narrows with time passing, and approaches a steady-state bandwidth of $\Delta\omega = a$ (3 dB cutoff frequency). After signal onset, the steady-state is reached at about $t = 1/a$ ($1/a$ is also the time constant of the exponential weighting). The damping factor a can be employed to control the steady-state bandwidth (that can be narrowed, however at the cost that the time needed to attain the steady-state proportionally increases). For simple cosine signals of sufficiently high frequency, the FTT magnitude spectrum according to

[5] The same consideration was made in "running" autocorrelation algorithms, which typically "slide" along a time signal and include a weighting function to successively discard past sample values so that ACF in fact is computed from an "effective time window" of N samples up to the sample point t moving with time. As to the equivalence of "running" ACF and FTT, see Terhardt 1998, 94f.

[6] A more detailed analytic formulation of the FTT is given by Mummert 1997.

Terhardt (1985, 254) is *largely similar to the output of a simple-resonance filter* for which the 3dB bandwidth is $B = a/\pi$. Given that the boundary between transient part and steady-state part can be taken as the "effective time window" of the analysis defined by $1/a$, the product of the effective time window and the steady-state bandwidth would be as small as $1/\pi = 0.3183$.

If this product would be viewed in terms of the uncertainty relation in regard to signals and systems, it would clearly be far below Gabor's theoretical limit of $\Delta f\, \Delta t = ½$. In this context, it might be noted that, for signals of given (rms) duration and energy (set to a value of 1), the uncertainty product has been calculated by Papoulis (1962, 62f., Equations 4-39 to 4-46) as

$$D_t * D_\omega \geq \sqrt{\frac{\pi}{2}} \tag{13}$$

where the equality holds for Gaussian signals (i.e., the product numerically yields 1.2533). The difference between products $\Delta f\, \Delta t \geq 1$ (Equ. 2) postulated from mathematical analysis and values much smaller than 1 calculated for FTT and other filter models results from the 3dB bandwidth parameter, which is common to filter design and performance tests yet must not necessarily apply to auditory perception. The bandwidth of the AF as determined in hearing experiments involving subjects of different age (Patterson et al. 1982) can be roughly given as 11% of the center frequency for young adults who have not yet suffered hearing loss. For a f_c of 0.5, 2 kHz and 4 kHz (as were employed in the experiments of Patterson et al. 1982), this means a relative filter bandwidth of ca. 191 cents (corresponding to the musical interval of a major second). Alternatively, the normalized width of the equivalent rectangular filter (roex[p, r]) has been given as $BW_{ER/fc} = 4/25 = 0.16$ (Patterson et al. 1982, 1801).

In FTT analysis, parameter values for bandwidth B and damping factor a can be set so as to simulate performance of the auditory periphery. To this end, the bandwidth should be that of the CB (cf. Zwicker & Fastl 1999, ch. 6) divided by 25, which would not be too far away from the jnd for pure tones[7]. Referring to analytical expressions designed to approximate critical-band rate and critical bandwidth (Zwicker & Terhardt 1980), Terhardt suggested that an "audio FTT" could be performed with the parameters set like

$$B = a/\pi = 1 + 3(1 + 1.4(f/\text{kHz})^2)^{0.69}\ \text{Hz} \tag{14}$$

Assuming that there are 24 CBs (expressed as a Bark scale), the frequency resolution for the FTT is 24 x 25 = 600 frequency samples per spectrum deemed sufficient and necessary to model peripheral auditory analysis (cf. Terhardt 1985, 255). In regard to the effective window length (i.e., the analysis interval T_A) relative to frequency bands, Terhardt (1992, 378) has given these figures:

f/kHz	0.1	0.5	1	2	4	8
T_A/ms	24	22	16	8	2.7	0.74

[7] For example, one CB included in the table given by Zwicker & Terhardt 1980, 152 ranges from 920 to 1080 Hz with f_c = 1000 Hz and is 160 Hz wide; divided by 25, the frequency step would be 160/25 = 6.4 Hz as compared to the jnd at 1ooo Hz, which is ca. 3 Hz.

Numerically, for a sampling rate at 44.1 kHz, an effective window length of 24 ms would correspond to 1058 samples falling into this time interval. A cosine signal of f = 0.1 kHz and a period of 10 ms would cover 441 samples per period so that the analysis interval will have access to, on the average (as the analysis window slides along the time signal), two periods of the signal. The ratio is much better at higher signal frequencies and shorter periods where the analysis window would hold (at best, if no truncation occurs) 16 periods at 1 k Hz as well as at 2 kHz. The effective window length of the FTT has been calculated (Vormann & Weber 1995, 1191) as

(15) $T(\omega) = 2.988/a(\omega)$

where $a(\omega)$ is the frequency-dependent transformation parameter. Correspondingly, the bandwidth is given as

(16) $$B(\omega) = \frac{\sqrt{\sqrt{2}-1}}{\pi} \cdot a(\omega)$$

whereby an uncertainty product T x $B \approx 0.61$ has been calculated. This of course would outperform a conventional Fourier transform analysis by far so that time/frequency resolution close to the cochlear filter bank can be expected from the FTT analysis (see below). In some of the relevant publications (Heldmann 1993, Vormann 1995), values as to T and B as well as to their product differ somewhat; parameter values as found in the literature for the 1st and 2nd order as well as estimates for the 4th order are given in table 1:

Table 1: FTT parameters

Order	1	2	4
Window function	e^-*ax*	x e^-*ax*	x³/6 e^-*ax*
Resolution *dT*	1 / *a*	2,988 / *a*	4,990 / *a*
Bandwidth (*B*)	*a* / π	0,6436 *a* / π	0,4350 *a* / π
dT * *B*	1 / π	1,923 / π	2,171 / π

In this table, a denotes the scaling factor $a(\omega)$, and t denotes the time axis. For practical reasons, parameter values may be rounded like

Order	1	2	4
Window function	e^{-at}	$t \cdot e^{-at}$	$\frac{t^3}{6} \cdot e^{-at}$
dT	1 / *a*	3 / *a*	5 / *a*
B	*a* / π	0,644 *a* / π	0,435 *a* / π
dT * *B*	1 / π ≈0,32	1,93 / π ≈0,61	2,17 / π ≈0,69

The bandwidth B for any order of analysis n can be calculated according to

$$(17) \qquad B = \frac{a}{\pi}\sqrt{2^{\frac{1}{n}} - 1}$$

The original FTT algorithm (see Terhardt 1985) has been improved later on in regard to the weighting function (cf. Schlang & Mummert 1990, Terhardt 1998, 97) where a form $a\ t\ e^{-at}$ has been proposed. Also, weighting of the form $h(t) = t^3 e^{-at}$ has been introduced for a 4th order FTT (as $h(t)$ in this case is equivalent to the Laplace transform of a 4th order low-pass filter, see von Rücker 1997).

For comparison of conventional Fourier transform and FTT analysis, a number of natural sounds were chosen; in addition some complex sounds based on FM and AM processes were generated with Mathematica. In the following, the results for the organ sound (Quintadena 16', pipe/note C_2) on which a bell sound has been superimposed (see figs. 4, 5, 6) will be presented.

In the FTT algorithm applied to analysis, a 4th order weighting function had been implemented. Since the effective time window for the standard FTT has been given as 24 ms at 0.1 kHz, corresponding to 1058 samples at 44.1 kHz sampling (see above), a comparison to an FFT of 1024 sample points seems a reasonable choice. However, the FFT also employed a weighting function for which a Blackman window was chosen[8].

The analysis obtained with a FFT of 1024 and Blackman weighting is shown in figure 10:

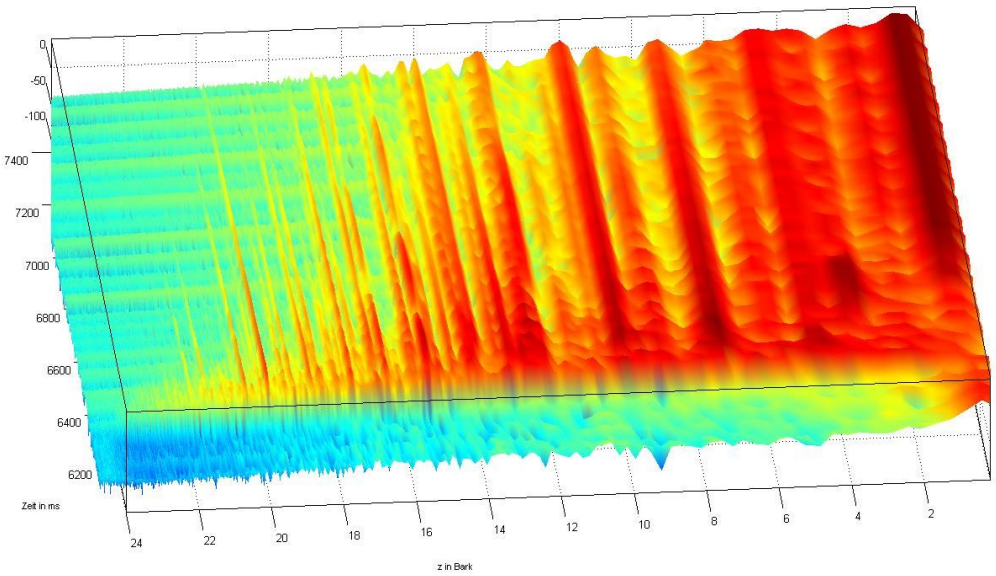


Figure 10: organ (Quintadena 16' C_2) plus bell, FFT 1024 pts, Blackman

[8] The ENBW for the Blackman window is 1.73 bins in DFT and the 3.0dB bandwidth is 1.68 bins.

The same sound subjected to 4th order FTT analysis is displayed in figure 11:

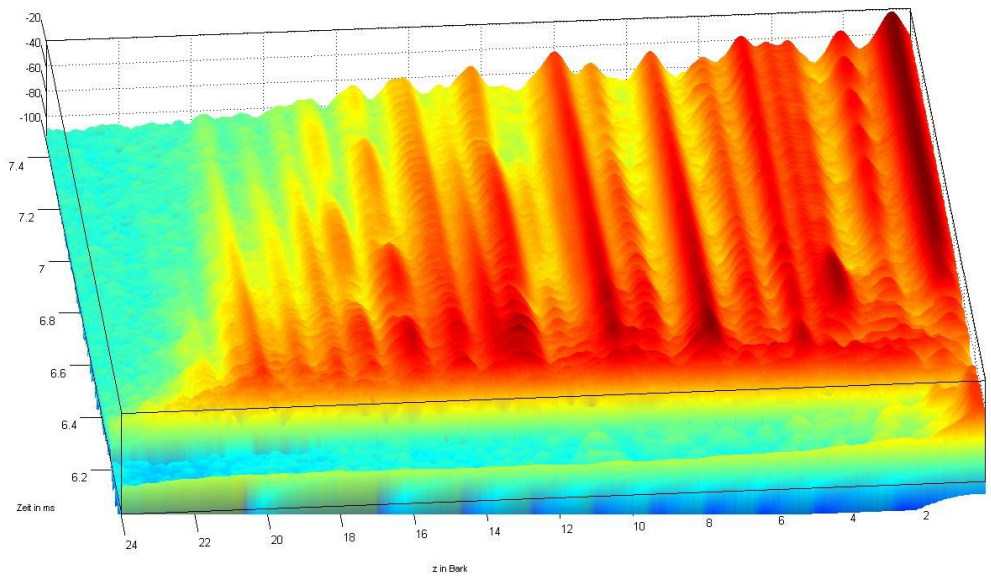


Figure 11: organ (Quintadena 16' C_2) plus bell, 4th order FTT

From a comparison of both analyses presented as 3D-plots (were the abscissa [x] is in Bark[z], the ordinate [y] is in dB, and time (ms) is in the z-dimension) one can see that time and frequency resolution for the FTT at low frequencies is considerably better than with the 1024 point FFT subjected to Blackman weighting. Note that with a FFT length of $N = 1024$ and sampling at 44.1 kHz, frequency resolution (Equ. 3) nominally is ca. 43 Hz. As this is the constant bandwidth of the FFT analysis (a DFT can be viewed as equivalent to a filter bank), the signal is under a fine-grain analysis at higher frequencies (Bark[z] 10-20) so that the FFT analysis picks many small spectral components corresponding to higher modes of vibration of the bell while the FTT analysis is more condensed since it relates to the concept of CBs, and therefore integrates such components which are closely spaced in frequency into broader "spectral ridges" (figure 10). A similar picture would be obtained with a WT-based analysis. One can argue that auditory perception of complex sounds basically is directed at picking spectral peaks that are present during a reasonable time interval (relevant as 'integration constant' in regard to hearing). In this respect, a limited number of clearly expressed "spectral ridges" may be more relevant to actual hearing as this must be performed in quasi-real time, and consequently calls for some temporal as well as spectral integration (as reflected in CBs and 'integration constants'). Algorithms directed to finding peaks in spectral envelopes are quite common as in LPC (see fig. 4) or similar source-filter analysis models (cf. Rodet & Schwarz 2007); if a sequence of frames is processed so that spectral envelope peaks can be separated and extracted, the next step is to connect such peaks from one frame to the next so that 'tracks' for harmonic partials or inharmonic components result over time. Such tracks then can be used for finding quasi-continuous pitch contours or for separation of 'sound objects' in a computational auditory scene approach (cf. Kostek 2005).

Comparison of the two types of analysis ("plain" Fourier, FTT) may indicate an advantage on the side of the FTT as one would expect from uncertainty products reported in the literature. However, the difference obtained in several analyses (of which but one example is included in the present article) seems gradual rather than principal. To optimize analysis, one often has to experiment with parameter settings. In addition, it is always revealing to apply different methods and models to the analysis of particular sound samples because in this way one can try to extract as many distinctive features as is needed for a certain problem, and at the same time the results obtained with one method can be tested for validity and reliability by using a second or even a third tool.

As far as 'perceptually adequate' analysis is concerned, comparison of several models including Gabor filtering, a linear, simplified but functional cochlear model (first published by Netten & Duifhuis 1983), WT and gammatone filtering tested for their impulse responses resulted in kind of a ranking (Hut, Boone, Gisolf 2006) where Gabor filtering was leading in regard to the uncertainty product, but also the linear cochlea model performed well. WT was judged to be unsuited to auditory modeling because an 'auditory wavelet' would not exist, and, therefore, Hut et al. (2006, 633) concluded that *wavelet analysis methods cannot be used in perception research*. The gammatone filter (implemented in many auditory models) according to these tests did well in terms of *general purpose linear time-frequency filtering, but does not give a good cochlear representation* (Hut et al. 2006, 635). Since an advanced cochlear model (Mammano & Nobili 1993, Nobili & Mammano 1996) seems to provide extremely good resolution in both time and frequency (Russo, Rožić, Stella 2011) with $\Delta f \, \Delta t \approx 0.55$, and hence close to the Gabor limit of 0.5, this approach perhaps could be the most promising to approximate performance of the auditory system even further (for recent developments, see Meddis et al. 2010). It should be noted, in this respect, that known values for the 'uncertainty relation' have been questioned to hold for the human auditory system (see, e.g. Kral & Majérnik 1996). The reason for such an assessment based on empirical data in most cases was that the performance of the auditory system in discrimination tasks (where stimuli were varied in frequency, level, and duration) was better than accepted values for the 'uncertainty product', on the one hand, and the relation between bandwidth and duration apparently was not linear, on the other. An explanation for this system behaviour can be found on the level of functional neuroanatomy and neurophysiology since hearing is effected by a complex network involving ascending and descending pathways as well as feedback regulation loops (as in OHC motility and BM/TM adjustment necessary for sharp frequency discrimination and 'pitch' processing; OHC = outer hair cell, BM = basilar membrane, TM = tectorial membrane; see Pickles 2008).

5. Conclusion

The present article intends to shed light on several approaches to digital sound analysis that are viewed (a) as tools useful for research in musical acoustics and organology, and (b) in regard to auditory perception. Besides the proven Fourier analysis techniques such as STFT, especially for the study of transient or impulsive sounds other methods such as WT (see Zhu & Kim 2006) or AR can be applied for time/frequency representations. To account for characteristics of the auditory systems, namely different resolution power relative to the period length (ms) of nearly periodic as well as quasi-periodic sound signals (meaning

spectral structures ranging from harmonic to inharmonic; see Schneider 1997, 2001), algorithms simulating peripheral filtering must be designed which offer appropriate filter bandwidth and time constants. WT and gammatone filter banks are among such algorithms that can be applied to many sounds, and can thus be considered versatile tools. If an approach is needed which is closer to functions found implemented in the auditory system, computational models such as developed by Meddis and O'Mard (1997, 2006) should be applied to the study of musical sound in regard to psychoacoustics and perception (see Schneider & Frieler 2009). The FTT model that was proposed already in 1985 still can be a useful method for time/frequency analysis that is close to basic parameters of the auditory periphery.

References

Bachmann, Werner 1992. *Signalanalyse. Grundlagen und mathematische Verfahren*. Braunschweig: Vieweg.

Beauchamp, James 2007. Analysis and Synthesis of musical instrument sounds. In J. Beauchamp (ed.). *Analysis, Synthesis, and Perception of musical sounds*. New York: Springer, 1-89.

Bilsen, Frans, Its Kievits 1989. The Minimum Integration Time of the Auditory System. *Preprint 2746, AES Convention Hamburg March 1989.*

Boersma, Paul 1993. Accurate short-term Analysis of the fundamental frequency and the harmonic-to-noise ratio of a sampled sound. *Proc. Institute of Phonetics, Univ. of Amsterdam* 17, 97-110.

Bracewell, Ronald 1978. *Fourier Transform*. 2nd ed. New York: McGraw-Hill.

Bregman, Albert 1990. *Auditory Scene analysis*. Cambridge, MA: MIT Pr.

Bürck, W., P.Kotowski, H. Lichte 1935. Der Aufbau des Tonhöhenbewußtseins. *Elektrische Nachrichtentechnik* 12, 326-333.

Cohen, Leon 1995. *Time-Frequency Analysis*. Upper Saddle River, N.J.: Prentice – Hall.

de Boer, Egbert 1976. On the "Residue" and Auditory Pitch Perception. In W.D. Keidel, W. D. Neff (eds.). *Handbook of Sensory Physiology* Vol. V, 3. Berlin, New York: Springer, 479-583.

de Cheveigné, Alain 2005. Pitch Perception Models. In Chr. Plack, A. Oxenham, R. Fay, A.Popper (eds.). *Pitch. Neural Coding and Perception*. New York: Springer, 169-230.

Dellomo, Michael, Garry Jacyna 1991. Wigner transforms, Gabor coefficients, and Weyl-Heisenberg wavelets. *Journ. Acoust. Soc. Am*. 89, 2355-2361.

Dutilleux, Pierre, A. Grossmann, Richard Kronland-Martinet 1988. Application of the Wavelet Transform to the Analysis, Transformation and Synthesis of musical Sound. *Preprint 2727*, *AES Convention 85*, Nov. 1988.

Eddins, David, David Green 1995. Temporal Integration and Temporal Resolution. In B.C.J. Moore (ed.). *Hearing*. San Diego etc.: Academic Pr., 207-242.

Evangelista, Gianpaolo 1997. Wavelet representations of musical signals. In C. Roads, St. Pope, A. Piccialli, G. de Poli (eds.). *Musical Signal Processing*. Lisse etc.: Swets & Zeitlinger, 127-153.

Flandrin, Patrick 1999. *Time-frequency/Time-Scale Analysis*. San Diego etc.: Academic Pr.

Gabor, Dennis 1946. Theory of Communication. *Journ. Inst. El. Eng*. 93, 429-457.

Gafori, Franchino 1496/1967/1968. *Practica musicae*. Milan (Reprint Farnborough, Hants.: Gregg Pr. 1967); (English translation and transcription of musical examples by Clement Miller; [no place]: Am. Inst. of Musicol. 1968).

Greenwood, Donald 1990. A cochlear frequency-position function for several species – 29 years later. *Journ. Acoust. Soc. Am*. 87, 2592-2605.

Heldmann, Klaus 1993. *Wahrnehmung, gehörgerechte Analyse und Merkmalsextraktion technischer Schalle*. Ph.D. thesis, Technical Univ. Munich.

Hut, Rolf, Marinus Boone, Andries Gisolf 2006. Cochlear Modeling as Time-Frequency Analysis Tool. *Acustica* 92, 629-636.

Jurado, Carlos, Brian Moore 2010. Frequency selectivity for frequencies below 100 Hz: comparison with mid-frequencies. *Journ. Acoust. Soc. Am*. 128, 3585-3596.

Keiler, Florian, Can Karadogan, Udo Zölzer, Albrecht Schneider 2003. Analysis of transient musical sounds by auto-regressive modeling. *Proc. of the 6th Intern. Conference on Digital Audio Effects* (DAFx-03), London: St. Marys, 301-304.

Kostek, Božena 2005. *Perception-based Data processing in Acoustics*. Berlin: Springer.

Kral, Andrej, Vladimir Majérnik 1996. Neural Networks simulating the frequency discrimination of hearing for non-stationary short tone stimuli. *Biol. Cybern*. 74, 359-366.

Küpfmüller, Kurt 1968. *Die Systemtheorie der elektrischen Nachrichtenübertragung*. 3rd ed. Stuttgart: Hirzel.

Mammano, Fabio, R. Nobili 1993. Biophysics of the cochlea: linear approximation. *Journ. Acoust. Soc. Am*. 93, 3320-3332.

Markel, John, Augustine Gray 1976. *Linear Prediction of Speech*. Berlin: Springer.

Marple, S. Lawrence 1987. *Digital Spectral Analysis*. Englewood cliffs, N.J.: Prentice-Hall.

Meddis, Ray, Lowel O'Mard 1997. A unitary Model of pitch perception. *Journ. Acoust. Soc. Am*. 102, 1811-1820.

Meddis, Ray, Lowel O'Mard 2006. Virtual pitch in a computational physiological model. *Journ. Acoust. Soc. Am*. 120, 3861-3869.

Meddis, Ray, Enrique Lopez-Poveda 2010. Auditory Periphery: from Pinna to Auditory Nerve. In R. Meddis et al. 2010, 7-38.

Meddis, Ray, Enrique Lopez-Poveda, Richard Fay, Arthur Popper (eds.) 2010. *Computational Models of the Auditory System*. New York: Springer.

Messner, Gerald 2011. Du krächzt wie ein Rabe…, singst wie eine Nachtigall…In A. Schmidhofer, St. Jena (eds.). *Klangfarbe*. *Vergleichend-systematische und musikhistorische Perspektiven*. Frankfurt/M.: P. Lang, 205-217 (plus sound examples on a CD in the book).

Mertins, Alfred 1996. *Signaltheorie*. Stuttgart: Teubner.

Mertins, Alfred 1999. *Signal Analysis*. Chichester: Wiley.

Meyer, Erwin, Dieter Guicking 1974. *Schwingungslehre*. Braunschweig: Vieweg.

Moore, Brian 1995. Frequency Analysis and Masking. In B. Moore (ed.). *Hearing*. San Diego etc.: Academic Pr., 161-205.

Moore, Brian 2008. *An Introduction to the Psychology of Hearing*. 5th ed. Bingley: Emerald.

Mummert, Markus 1997. *Sprachcodierung durch Konturierung eines gehörangepaßten Spektrogramms und ihre Anwendung zur Datenreduktion*. Ph.D. thesis, Technical Univ. Munich.

Nobili, R., Fabio Mammano 1999. Biophysics of the cochlea II: Stationary nonlinear phenomenology. *Journ. Acoust. Soc. Am*. 99, 2244-2255.

Oertel, Dora, Richard Fay, Arthur Popper (eds.) 2002. *Integrative Functions in the Mammalian Auditory Pathway*. New York: Springer.

Papoulis, Athanasios 1962. *The Fourier Integral and its applications*. New York: McGraw-Hill.

Patterson, Roy, I. Nimmo-Smith, D. Weber, R. Milroy 1982. The deterioration of hearing with age: frequency selectivity, the critical ratio, the audiogram, and speech threshold. *Journ. Acoust. Soc. Am*. 72, 1788-1803.

Patterson, Roy, K. Robinson, J. Holdsworth, D. McMcKeown, C. Zhang, M. Allerhand 1992. Complex Sounds and Auditory Images. *Advances in the Biosciences* 83, 429-443.

Pickles, James 2008. *An Introduction the Physiology of Hearing*. 3rd ed. Bingley: Emerald.

Pressnitzer, Daniel, Roy Patterson, Katrin Krumbholz 2001. The lower limit of melodic pitch. *Journ. Acoust. Soc. Am*. 109, 2074-2084.

Rodet, Xavier, Diemo Schwarz 2007. Spectral Envelopes and additive+residual analysis/synthesis. In J. Beauchamp (ed.). *Analysis, Synthesis, and Perception of musical sounds*. New York: Springer, 174-227.

Rossing, Thomas 1982. *The Science of Sound*. Menlo Park, CA: Addison – Wesley.

Rücker, Claus von 1997. Berechnung von Erregungsverteilungen aus FTT-Spektren. *Fortschritte der Akustik – DAGA 1997*, 484-485.

Russo, Mladen, Nikola Rožić, Maja Stella 2011. Biophysical Cochlear Model: Time-frequency analysis and signal reconstruction. *Acustica* 97, 632-640.

Schlang, M, Mummert, M., 1990: Die Bedeutung der Fensterfunktion für die Fourier-t-Transformation als gehörgerechte Spektralanalyse. *Fortschritte der Akustik, DAGA '90*, Bad Honnef 1990, 1043-1046.

Schneider, Albrecht 1997. *Tonhöhe, Skala, Klang. Akustische, tonometrische und psychoakustische Studien auf vergleichender Grundlage*. Bonn: Orpheus-Verlag für Syst. Musikwiss.

Schneider, Albrecht 2001. Complex inharmonic sounds, perceptual ambiguity, and musical imagery. In R.I. Godøy, H. Jørgensen (eds.) *Musical Imagery*. Lisse etc.: Swets & Zeitlinger, 95-116.

Schneider, Albrecht, Klaus Frieler 2009. Perception of harmonic and inharmonic sounds: results from ear models. In S. Ystad, R. Kronland-Martinet, K. Jensen (eds.). *Computer Music Modeling and Retrieval. Genesis of meaning in sound and music*. Berlin: Springer, 18-4

Schneider, Albrecht, Arne von Ruschkowski, Rolf Bader 2009. Klangliche Rauhigkeit, ihre Wahrnehmung und Messung. In R. Bader (ed.). *Musical Acoustics, Neurocognition and Psychology of Music*. Frankfurt/M.: P. Lang, 103-148.

Schneider, Albrecht, Valeri Tsatsishvili 2011. Perception of musical intervals at very low frequencies: some experimental findings. In A. Schneider, A. von Ruschkowski (eds.). *Systematic Musicology: Empirical and Theoretical Studies*. Frankfurt/M.: P. Lang, 99-125.

Solbach, Ludger, Rolf Wöhrmann, Jörg Kliewer 1998. The complex-valued continuous wavelet transform as a preprocessor for auditory scene analysis. In D.F. Rosenthal, H.G. Okuno (eds.). *Computational Auditory Scene Analysis*. Mahwah, N.J.: Erlbaum, 273-292.

Snyder, Bob 2000. *Music and Memory*. Cambridge, MA: MIT Pr.

Terhardt, Ernst 1985. Fourier Transformation of time signals: conceptual revision. *Acustica* 57, 242-256.

Terhardt, Ernst 1992. From Speech to language: on auditory information processing. In M.E.H. Schouten (ed.). *The Auditory Processing of Speech. From sounds to words*. Berlin, New York: Mouton de Gruyter, 363-380.

Terhardt, Ernst 1998. *Akustische Kommunikation*. Berlin, New York: Springer.

Vormann, Mathias, 1995. *Psychoakustische Modellierung der virtuellen Tonhöhe.* Diploma thesis (Physics), Carl von Ossietzky Univ., Oldenburg.

Vormann, Mathias, Reinhard Weber 1995. Gehörgerechte Darstellung von instationären Umweltgeräuschen mittels Fourier-Time-Transformation (FTT). *Fortschritte der Akustik – DAGA 1995*, 1191-1194.

Winer, Jeffery, Christoph Schreiner (eds.) 2011. *The Auditory Cortex*. New York: Springer.

Yen, N. 1987. Time and frequency representation of acoustic signals by means of the Wigner distribution function: Implementation and interpretation. *Journ. Acoust. Soc. Am*. 81, 1841-1850.

Zhu, Xiangdong, Jay Kim 2006. Application of analytic wavelet transform to analysis of highly impulsive noises. *Journ. Sound Vibr*. 294, 841-855.

Zwicker, Eberhard, Ernst Terhardt 1980. Analytical expressions for critical-band rate and critical bandwidth. *Journ. Acoust. Soc. Am*. 68, 1523-1525.

Zwicker, Eberhard, Hugo Fastl 1999. *Psychoacoustics. Facts and models*. 2nd ed. Berlin: Springer.