ARCHIVES OF ACOUSTICS
24, 2, 191–206 (1999)

FROM EAR MODELING TO AUDITORY TRANSFORM

P. KLECZKOWSKI

Department of Mechanics and Vibroacoustics
Academy of Mining and Metallurgy
(30-059 Kraków, Al. Mickiewicza 30, Poland)
e-mail: kleczkow@uci.agh.edu.pl

Understanding how the auditory system works recently gains increasing importance in audio engineering. Its most widespread practical use is in perceptual audio coders, with even more applications to be foreseen in the future. The construction of a mathematical procedure that could transform acoustic signals heard by humans to data corresponding to auditory sensation would open way to significant progress in audio engineering. In this paper the issue is discussed and important research in the field is reviewed. The proposal for a frequency analysis procedure appropriate for ear modeling is presented and verified. This procedure is a form of the Wavelet Transform.

1. Introduction

Understanding how the ear works has been a challenge to researchers since Ohm formulated his "acoustic law" in 1843. This knowledge, apart from purely medical, has practical applications in such fields as construction of devices for people with handicapped ear, audio engineering [37] or noise control [11]. A new area of technology where ear modeling is directly applied is the construction of low bit rate coders of digital audio signals, usually referenced as perceptual coders.

Another, future-oriented application is the construction of an appropriate set of data (data vector) to be used as an input to artificial neural networks.

The foundations for modern knowledge on this subject have been laid by von Bekesy in 1947, with his pioneering work identifying vibration at the basilar membrane as travelling waves. The efforts are continued, but functioning of higher (neural) stages of the ear are still subjects of partial hypothesis, of which not one has been widely accepted without questioning.

2. The auditory system as an acoustic receiver

Our ear perceives time functions of acoustic pressure as continuous evolution, or sequences, of separate acoustic events, characterized by their pitches and timbres. It is able to perceive and comprehend speech of one speaker with many other voices in the background, or to follow the melodic line of one instrument out of an ensemble. The auditory system transforms a time function of acoustic pressure into separate streams of information. Most of these complex functions are performed in higher stages of the auditory system and little is known about them, but whatever the nature of this information-processing mechanism is, it must rely on information received from the peripheral stages of the auditory system. Therefore, any attempt to investigate these processes must be based on a solid model of information supplied by the peripheral auditory system to higher auditory stages.

The time function received by the ear is a linear superposition of acoustic pressures from separate individual sources, while the separate streams of information isolated by the ear build auditory sensations corresponding to sounds produced by each of the sources individually. Although the number of those individual sources that can be simultaneously perceived by the ear is limited, the entire mechanism is so sophisticated that no successful attempt has been made so far to implement it artificially.

3. General review of operations performed by the auditory system

In this paper the acoustical, the mechanical and a fragment of the neural path are considered. The last physiological (neural) element in this chain that is covered is cochlear nucleus, so little more than the peripheral auditory system and only monaural hearing will be discussed.

The outer ear plays roles of mechanical protection, microclimate control and directivity detection aid. Physically, it is a pipe with a length of about 2 cm open at one end thus enhancing waves of four time its length, i.e. frequencies around 4 kHz. The middle ear, i.e. three ossicles: malleus, incus and stapes act as an impedance transformer, which maximizes the energy transmitted into the inner ear, and eliminates reflections of waves at the boundary of gas and liquid mediums. The cochlea (part of the inner ear) is filled with a liquid (perilymph and endolymph), and because of complexity of mechanics of the cochlea, the input impedance of the inner ear changes with vibration level and frequency. At the oval window — the place where the stapes touches the cochlea and which is a boundary between the middle and inner ears, the vibration of the stapes can still be considered a time function, related to the acoustic pressure outside of the ear.

Inside the cochlea complex, but relatively well understood processes take place, which can be functionally (but hardly physiologically) separated into two sub-processes. The first is nonlinear mechanical filtering, the second is a transduction from mechanical movement of the basilar membrane to electrochemical activity of neurones.

The basilar membrane with perylimph surrounding it was long considered to be a passive mechanical filter. Vibration at the oval window is the source of waves travelling along the membrane up to the point of maximum amplitude, and then decaying rapidly. The basilar membrane can be considered as a bank of bandpass filters. As all filters have steep upper slope and mild lower slope simpler models often simulate it as a chain of low-pass filters. The estimation of the resolution of this mechanical filter has changed substantially. Earlier experiments indicated that the Q factor of a typical basilar membrane — based bandpass filter was on the order of 1, that is far lower than the overall Q value of the entire auditory system estimated at 100 (or even 200 by some researchers [39]). Over the past 25 years more sophisticated measurement methods were developed and evidences became available that the true value of Q of the membrane is much higher [28]. No passive mechanical model can explain such a high selectivity.

It is now agreed by most researchers that there is some active process in the cochlea, supplying energy to the basilar membrane in a positive feedback which provides this high frequency selectivity, and that outer hair cells in the organ of Corti, a small organ distributed along the basilar membrane, play a key role in this process.

The other sub-process performed in the cochlea is a transduction of vibration at specific place along the basilar membrane to the appropriate neural signal [34, 39]. This process is performed inside the organ of Corti, and transducers are inner hair cells. The vibration sensed is, in terms of signal processing, rectified, and then converted to series of neural spikes further transmitted along the auditory nerve (composed of around 30000 neural axons) to higher stages of the auditory system. The rectification mentioned is only approximate and has physiological origin: the processes resulting in the excitation of hair cells are in the most part those accompanying the deflection of the basilar membrane in one direction.

There is a long lasting controversy among researchers of the hearing system. One point of view attributes most meaning to the analysis of places of excitation on the basilar membrane ("place theories") while the other emphasizes the role of information contained in the time structure of a signal from a particular area on the basilar membrane ("time theories").

Nonlinearities in the operation of the cochlea have three sources. One is mechanical nonlinearity of the basilar membrane. The second is inherent nonlinearity of the electrochemical processes in mechanical to neural transduction. This latter nonlinearity is of the "hard limiting" type and is seriously limiting the dynamic range of any single hair cell. However, the entire transduction mechanism compensates for this, by way of combining many haircells with different thresholds of activation [28].

The third source of nonlinearity is the positive feedback in the cochlea mentioned above, and is least known.

Experiments aimed at estimation of the Q value at different stages of the neural path of the auditory system have shown that the shape of tuning of neural responses along the higher stages of the auditory system is similar to the response at the auditory nerve leaving the cochlea and is not very sharp. The origin of the ear's high Q factor is still not quite clear [28].

One feature, commonly agreed upon is that the representation of neural activity leaving the cochlea is tonotopic throughout the rest of the auditory system, that is spatial distribution of neural cells conveying information from particular fragments of the basilar membrane is preserved along the subsequent stages of the auditory system. There is evidence that this tonotopic distribution becomes two-dimensional in the auditory cortex [41]. The higher stages of auditory processing are fairly well known in their physiological construction, also neural responses in different stages have been intensively studied [28].

However, functional explanation of the entire system is still rather unclear. Apart from higher functions mentioned in Sec. 2. the seemingly simpler mechanism of a very high frequency selectivity of the ear has not yet been understood.

One more physiological stage of neural processing within the scope of this article is the cochlear nucleus. There are evidences that the process called lateral inhibition takes place there. When largely simplified and in the context of spectral analysis performed by the ear, it can be described as enhancing stronger spectral components while suppressing neighbouring weaker components. However, it is not clear whether this mechanism contributes to the sharpening of ear's overall frequency selectivity.

4. Different approaches to modeling the auditory system

This task is enormously complex for two main reasons:

a) The elements of the auditory system are highly nonlinear and difficult to separate into independent blocks performing specific functions. Modeling its higher stages is yet more difficult because their operation is hardly known.

b) When modeling the entire auditory system we have to deal with the output signals which are very difficult to measure since we neither have a measuring device nor a unit of measure. If such a unit existed, it would have to be multidimensional. Some works addressed the problem of multidimensional timbral space, trying to locate sounds of musical instruments there [33]. It is difficult to express subjective percepts in our brain in a quantitative way. Auditory percepts are the domain of psychoacoustics and some quantitative measures of auditory sensations have been developed there, for example the sone scale for measurement of sensation of loudness. Although psychoacoustics can help us to measure the sensations of loudness and pitch, we are still unable to measure more complex, multidimensional percepts such as timbre.

The problem of modeling can be approached by looking at it from two different perspectives.

Psychoacoustical approach. It seems that many of the test stimuli used in psychoacoustical experiments may not engage more complex functions of the higher stages of the auditory system, so that they may not reveal some features of the ear.

Physiological approach. Here an attempt is made to divide the auditory system to some physiologically separate parts and model their functions mathematically. A model encompassing all peripheral stages and higher stages up to the cochlear nucleus, with consequent mathematical formalism can be found in [39]. However, only limited verification of this model is given.

Fairly well verified models of this sort exist, but most of them encompass only limited part of the auditory system, mainly in its peripheral stages.

Some works try to model particular functions of the ear, instead of its physiological parts. Such models differ from the psychoacoustic approach in that in building these models they use knowledge about physiological construction of the system and can be seen as physiologically encompassing the complete system.

The function most often simulated in such models is the sensation of pitch [8, 12, 27, 36]. The sensation of pitch is very suitable for modeling, for the ear is very sensitive to pitch (high overall Q of the system seen as a filterbank) and it is probably the most important percept upon which the ear analyses and qualifies sounds.

5. Techniques of modeling

Probably the largest number of physiologically-based models have been built for the basilar membrane, from linear one-dimensional ones of various complexity [12, 39], non-linear one-dimensional [21], to three-dimensional [3]. In [20, 24] models in the form of chains of building blocks are presented. These implementations are analog and digital respectively, and both include positive feedback activated by outer hair cells.

When modeling the peripheral part of the ear, the different physical quantities (mechanical and electrochemical) can be modeled as continuous or discrete functions of time. Some researchers tend to exploit information theory and shift the problem from the domain of deterministic signals to the domain of stochastic signals, like in [13, 30].

Modeling of higher neural stages requires sophisticated mathematical tools used in pattern recognition, neural network modeling and probably some specific ones not yet developed.

6. Decomposition of functions of the ear into blocks

The following block diagram of the ear may be proposed (Fig. 1), if we try to use blocks to which appropriate signal processing procedures can be attributed This is a framework generalizing the approach taken in some models. Such a generalized model attempts to simulate all salient functions of peripheral stages of the ear, albeit in simplified form.

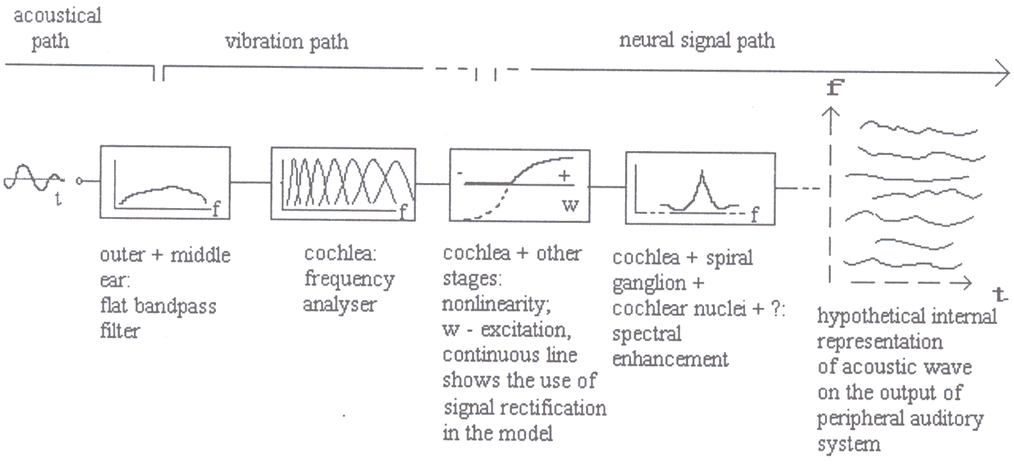
acoustical path
vibration path
neural signal path
t
f
f
+
-
w
f
f
t
outer + middle ear: flat bandpass filter
cochlea: frequency analyser
cochlea + other stages: nonlinearity; w - excitation, continuous line shows the use of signal rectification in the model
cochlea + spiral ganglion + cochlear nuclei + ?: spectral enhancement
hypothetical internal representation of acoustic wave on the output of peripheral auditory system

Fig. 1. Simplified functional diagram of extended peripheral auditory system.

A common element of all models of the ear is some sort of frequency analysis (Fig. 1), as this function of the cochlea is unquestionably agreed upon as probably the most fundamental. Many models in fact perform only this one function and still are successful, especially when frequency analysis is combined with a simplified form of the fourth block in Fig. 1 — spectral enhancement.

The nonlinearity, performed in the third block is being simulated by different means, and with very different effect on processing of the signal. However, due to its very complex nature and many places of the auditory system where it can be found, it is rather difficult to simulate its effects [28, 43]. Some models include the automatic gain control as the only nonlinearity, as such mechanism is certainly present in the auditory system. This stage will not be discussed in more detail.

Spectral enhancement (fourth block) is associated with the reduction of data. Such a process takes place in the auditory system. A simplified implementation of this block is achieved by assigning less meaning to parts of the spectrum which are weaker. Less meaning leads to less bits of resolution and the procedure is governed by masking curves supplied by psychoacoustics. This is the way all perceptual coders work.

7. Modeling the frequency analysis function of the ear

Bandpass filters appropriate for modeling the ear should have shorter impulse response for high frequencies than for low ones [17, 19, 31]. Such are the "filters" in the ear. This type of filters is usually referred as "constant Q" type, i.e. the widths of bandpass filters are proportional to their centre frequencies.

For frequency analysis typically either the spectral decomposition by means of an appropriate transform, or a bank of bandpass filters is used. Throughout this paper, the formulas will be given either in continuous or in discrete form, whichever is more convenient.

Usually a sequence of block transforms:

$$\mathbf{y}_k = \mathbf{H}\,\mathbf{x}_k, \tag{1}$$

where $\mathbf{y}_k$ is k-th consecutive output vector, $\mathbf{x}_k$ is k-th consecutive input signal vector and $\mathbf{H}$ is a transform matrix (time-invariant) is used, to obtain a time — frequency representation of the signal. Then the filtering approach is closely related. In their basic forms both techniques can be shown to be fully equivalent [40]. However, differences in implementations are meaningful. The comparison of both approaches, considered as candidates for modeling of auditory spectral analysis, leads to the following conclusions.

The transform approach:

+ compact mathematical form, often resulting in powerful fast computing algorithms,

+ the transformed signal has a form readily available for monitoring, either for a human observer or for an algorithm simulating further functions of the ear,

+ orthogonal transforms obey the Parseval's theorem,

– only wavelet transforms have a very desirable feature of performing a constant Q type of spectral analysis, instead of most common constant bandwidth type,

– distortion may be generated in boundaries of neighbouring signal blocks (blocking effect), which requires means to alleviate or eliminate this problem.

The filtering approach:

+ is more natural in simulation of auditory filters,

+ arbitrary filter shapes can be designed, especially in the case when non-decimated (operating at the sampling frequency of the input signal in all bands) filters are used,

+ has no blocking effects,

– non-decimated filters lead to very redundant spectral representation,

– decimated filters (critically sampled) can only be used when none or very limited processing in spectral domain is to be performed, as those filters guarantee perfect reconstruction only under such a condition,

– if some distortion is present resulting from imperfections of filters, they are very different in nature than those produced by the ear.

7.1. Application of transforms

The Fourier Transform in its original form can only be used for either the analysis of impulses shorter than ear's resolution in time or for stationary signals, and hence is impractical for ear modeling, where we require a representation in which the output vector $\mathbf{y}_k$ changes with time, as in (1). Such a distribution is the Short Time Fourier Transform (STFT — the definition in continuous domains is given as more general) [29]:

$$\mathrm{STFT}\,(\tau,\omega) = \int\limits_{-\infty}^{\infty} s(t)h(t-\tau)e^{-j\omega t}\,dt, \tag{2}$$

where τ is time around which we analyse the spectrum, ω is radian frequency, $s(t)$ is an input signal, $h(t)$ is time window through which we observe this signal. The shape of the time window determines the parameters of the distribution obtained. It is easy to notice that if we divide the integrand by $e^{-j\omega t}$ then we obtain the convolution of a time signal $s(t)$ with $h(-t)$, i.e. the time window reversed in time. This is where equivalence between a transform and a filter comes from. The Fourier transform $H(\omega)$ of the function $h(t)$ is exactly the shape of the filter equivalent to the evolution with time τ of one point of the transform.

The STFT as a tool for ear modeling has two drawbacks. The first one is that it performs a constant bandwidth analysis. The second results from the Balian–Low theorem [23], which states that if we critically sample the STFT (to avoid redundancy) then either time or frequency support of basis functions of the transform must go to infinity, thus good time and frequency localization is not possible.

More advanced transforms, related to the STFT have been developed and are used in practical implementations of perceptual audio coders. They are easily interpreted as filters. Most popular of them is the Modulated Lapped Transform also known as Modified Discrete Cosine Transform (MDCT) [25, 35]. The more general framework that describes them is called local trigonometric bases [44]. Being orthogonal transforms, in contrast to the STFT they can achieve good localization in both time and frequency. They also meet a specific compromise between the Q factor of filters, their distortion rate and computational efficiency, and would be very good candidates for ear modelling, but they perform constant bandwidth analysis.

Linear transforms related to the Fourier transform are all limited in their time--frequency resolution by the uncertainty principle. This limit could be relaxed by the use of one of a family of time-frequency distributions which have much better energy concentration in the time-frequency plane (Cohen's class distributions) [7]. The most widely investigated of them is the Wigner–Ville distribution:

$$W(\tau,\omega)=\int\limits_{-\infty}^{\infty} s\left(\tau+\frac{t}{2}\right)s^{*}\left(\tau-\frac{t}{2}\right)e^{-ej\omega t}, \tag{3}$$

where $s^{*}(t)$ denotes complex conjugate of the signal. However, as can be seen from the formula this representation is quadratic in $s(t)$ and therefore the distribution is nonlinear. It has the so-called "cross-terms" which make the results quite obscure, despite their excellent time-frequency resolution. It is also, in general case, non-invertible.

The cross-terms can be smoothed to some extent, but at the cost of reducing good joint time-frequency resolution.

Some works have addressed the problem of finding a frequency analysis tool appropriate to analyze sounds of musical instruments. Although not directly referring to the ear, their results are suitable for the problem discussed in this paper. A specific distribution belonging to Cohen's class, related to the Wigner distribution and called modal distribution is given in [32]. This distribution suppresses cross-terms but is inappropriate for transient — like sounds, such as musical sounds with sharp attacks.

There have been many attempts to modify the Fourier or related ransforms in order to make it a constant Q analysis. For the purpose of analysing musical sounds Brown [4] proposed such a modification directly in the digital form, but it was not invertible.

The relatively new transform, the Wavelet Transform (WT), is of the constant Q type, when used for frequency analysis [6, 10, 23, 38, 40]. The continuous WT(CWT) is given by:

$$\mathrm{CWT}(\tau,a)=\frac{1}{\sqrt{a}}\int s(t)h^{*}\left(\frac{t-\tau}{a}\right)dt, \tag{4}$$

where a is a scale factor (corresponding to frequency), $h(t)$ is a basic wavelet (or mother wavelet). The $h(t)$ can be real (more often used in practice) or complex; in this latter case the complex conjugate $h^{*}(\cdot)$ is used in (4) and the transform becomes complex. The choices of $h(t)$, and discretization steps for a and τ (together forming a discretisation grid in the time — scale plane) determine the features of the wavelet transform obtained.

The essence of the WT is that for any particular scale a the function $h(\cdot)$ of (4) is dilated or contracted proportionally along the abscissa, thus fulfilling the postulate of constant Q representation.

Generally, there can be four forms of the WT: the continuous WT given by (4); the discrete parameter WT, where parameters a and τ are discretized; the discrete time WT, where parameters and the signal under analysis $s(k)$ is discrete; the discrete WT, where the previous quantities and the $h(a,\tau,k)$ are discrete. A discrete mother wavelet

$h(k)$ can be a sampled version of its continuous time counterpart, but $h(k)$ which do not have a continuous version can be constructed. The latest case, with the basis value of a equal 2, takes the form:

$$\mathrm{DWT}(m,n) = 2^{-m/2} \sum_{k} s(k)h(2^{-m}k - n), \tag{5}$$

where m, n are integer values of parameters a and τ, k is a number of signal sample. This form is most widely used in practice, because of two important advantages:

a) the scaling forms a simple "dyadic" structure, for which fast algorithms — Fast Wavelet Transform exist;

b) the interpretation as a filterbank is straightforward.

The big disadvantage, when used for modeling of the ear, is that the analysis defined by (5) is an octave band analysis, so its frequency resolution is insufficient.

If $\{h_{m,n}\}$ form an orthonormal basis set, then the inversion formula takes the simple form of the sum of appropriate inner products between transform coefficients and basis functions:

$$s(k) = \sum_{m} \sum_{n} 2^{-m/2} h(2^{-m}k - n)\mathrm{DWT}_{m,n}. \tag{6}$$

The mathematics related to the choice of the mother wavelet is fairly complex [10], however, if we can accept large redundancy in our data, which in practice means choosing a relatively dense sampling of the time-scale plane, then we have large area of freedom in this choice.

In the redundant case the basis vectors of the tranformation will not be orthogonal, the only requirement is that they must span the vector space. The set of basis vectors is then called a frame. A condition for a set to be a frame is that for any $m \times 1$ vector $\mathbf{s}$ [6, 10, 23, 38]:

$$A||\mathbf{s}||^2 \le \sum_{i=1}^{n} |\langle \mathbf{s}, \mathbf{h}_i \rangle|^2 \le B||\mathbf{s}||^2, \qquad n \ge m, \tag{7}$$

where A and B are positive constants ("frame bounds"), $\mathbf{s}$ is input signal vector, $\langle \mathbf{s}, \mathbf{h} \rangle$ is the inner product of vectors. According to (7), the energy of the (discrete) wavelet coefficients relative to that of the signal must be within the two bounds.

The reconstruction from a frame is more difficult, as instead of the set $\{h_i\}$ another set, called dual set $\{\hat{h}_i\}$ is used [10], and its computation from $\{h_i\}$ is not easy. When in (7) $A = B$, then the frame is called tight frame and $\{\hat{h}_i\} = \{h_i\}$ holds.

When the ratio of B/A is close to 1, then a frame is an approximation to the tight frame, and signals can be approximately reconstructed with the use of an original basis set $\{h_i\}$. It depends on the application and the particular mother wavelet, how close to 1 the B/A ratio should be.

Working with the discrete parameter WT, we can obtain frames arbitrarily close to tight frames, by dense sampling of the CWT in both time and scale, i.e. the a and τ parameters. For a given mother wavelet there are threshold values (a_0, τ_0), below which the $\{h_{m,n}\}$ will always form a frame [6].

The fact that it is possible, albeit not easy, to use bases which are not orthogonal and thus to have large flexibility in the choice of a mother wavelet is crucial in the task of ear modeling: it allows to construct $h(t)$ which simulates the impulse response of auditory filters.

Irino and Kawahara [18] used a simulated response of the auditory filter as the mother wavelet, and called it the Auditory Wavelet Transform. As this resulted in a non-orthogonal transform, the reconstruction (as mentioned in Sec. 7.1) was not easy, and they used two indirect methods.

7.2. Application of filters

Besides the traditional classification of digital filters to FIR and IIR types, modern filters could be divided to non-decimated (standard) and decimated (usually critically sampled) filters. While critically sampled filters do not increase the data rate of the input signal they impose several important practical limitations. The construction of banks emulating constant Q filters is possible with the use of wavelet packets [44] and excellent filterbanks for the use in audio technology have recently been designed [2]. However, they tolerate little modifications to the channel signals, which are needed in ear modelling, e.g. in blocks no. 3. and no. 4. of Fig. 1.

Non-decimated filters offer much better flexibility, computational stability and robustness to distortion, and are thus more suitable for ear modelling. The cost of non-decimating is very high redundancy of the channel (output) data. Constant Q, one-third octave band or so-called Bark filters (with bandwith corresponding to 1 Bark) are often used [37].

Non-decimated filters with impulse responses simulating that of the cochlea have been investigated. In several models the so called "gammatone" linear filter has been used. Recently Irino [17, 19] has proposed a more advanced filter called "the gammachirp" with level — dependent, asymmetric characteristics, showing that it was theoretically optimum filter, with minimum uncertainty in a joint time frequency representation. It is given by [17]:

$$g_c(t) = at^{n-1} \exp(-2\pi b \,\mathrm{ERB}\,(f_r)t) \cos(2\pi f_r t + c \ln t + \phi), \qquad t > 0, \tag{8}$$

where a, b, c, n, are parameters, which are tuned to obtain best results, f_r is the centre frequency of the filter, ϕ is the phase of the cosine carrier, ERB is the Equivalent Rectangular Bandwidth and t is time. As the ERB is a function of f_r the bandwidths of cochlear particular filters can be precisely tuned to experimental data, thus offering better fit that the constant Q filters do.

Without the frequency modulation "chirp" term ($c \ln t$) the impulse response in (8) is equivalent to the earlier "gammatone" filter.

8. Spectral enhancement

Since a seminal paper by McAulay and Quatieri [26] several procedures for discarding weaker parts of the spectrum have been proposed. Such an operation has its foundations in physiological phenomena of masking and lateral inhibition (mentioned in

Sec. 3). It can be assumed that they both contribute at peripheral auditory stages to the overall ear's ability to isolate signals of interest out of an acoustic background. The procedures in [9, 16, 26] and many others are peak picking algorithms. They consist of finding local spectral peaks in subsequent spectra evolving with time and then in forming continuous traces in the three-dimensional amplitude versus time and frequency space.

On their output they all produce plots resembling the rightmost diagram in Fig. 1. Usually they provide substantial amount of data reduction. They differ in time-frequency analysis method used, selection (peak-picking) method, elimination of blocking effects and other details.

A simple way of achieving the same goal, albeit applicable only for stationary sounds of acoustic instruments with harmonic spectra was proposed in [22].

A different approach to modeling spectral enhancement was presented in [41, 42, 43]. The authors tried to simulate more precisely the physiological processes, with mixed partial derivative with respect to both time and space of the basilar membrane patterns as key contrast — enhancing operation, followed by nonlinearity (rectification).

9. Postulates for Auditory Transform

Auditory Transform is meant to be a multi-stage computational procedure (the term "transform" is used in a wide sense) which should serve two goals:

a) analyse an audio signal yielding a result similar to its internal representation at a suitable stage of the auditory system;

b) model the peripheral auditory system.

The most desirable properties of Auditory Transform are proposed below:

a) The first stage should be linear and perceptually invertible, i.e. a listener should not perceive any difference between an original signal and the reconstructed signal.

b) The careful design of frequency analysis part of this procedure is essential. If the WT is used for this stage, a good candidate is based on a mother wavelet that in some way approximates the characteristics of an auditory filter.

c) The subsequent nonlinear spectral enhancement procedure should eliminate redundancy from frequency analysis stage.

10. A proposal for a narrow-band wavelet transform

Following the discussion presented above, the author designed a specific wavelet transform to model the frequency analysis function of the ear. The choice of wavelet leads to a redundant (frame-based) wavelet transform. The construction and verification of the proposed WT is only summarized in this chapter.

There is a way of densely sampling of the CWT, keeping a simple structure of time allocation of transform coefficients. This is obtained by preserving a dyadic structure of sampling in time, with appropriate oversampling, while samples in the scale (frequency) domain are taken by filling the dyadic scale with additional samples, taken at fractional powers of two. Such samples in the scale domain are called "voices". If we denote the continuous wavelet at scale (octave) m by $h_m(t)$, then the voice number j in that scale will be given by:

$$h_{m,j}(t) = 2^{-j/M} h_m(2^{-j/M} t), \tag{9}$$

where M is the number of voices per one octave.

The mother wavelet chosen for that transform was similar to the Morlet wavelet. The original Morlet wavelet [6] is the basis function of the Gabor Transform, i.e. it is the complex sinusoid windowed by the Gaussian envelope. Its important advantage is that it directly preserves phase, as the basis functions are complex. The frequency resolution of the original Morlet wavelet was found inappropriate and a specific window was proposed instead of the Gaussian. This window was derived from the modified Blackman window [14], by tuning its coefficients so that the spectrum of the window is as close as possible to frequency characteristics of auditory filters, within limitations of real-valued windows. The discrete formula for the window was the following:

$$w(n) = 0.4205{+}0.4995 \cos\left[\frac{2\pi}{N}\right] {+}0.08 \cos\left[\frac{2\pi}{N}2n\right], \qquad n = -\frac{N}{2}, \dots, 0, \dots, \frac{N}{2}, \tag{10}$$

where $N + 1$ is the length of the window in samples. Odd length was used following the practice used in FIR filter implementations, where it is usually better to have the filter response sampled exactly in its center. The amplitude spectrum of this window is shown in Fig. 2. The frequency resolution of the resulting wavelet is much better that than offered by responses of filters described in [17] and [19].

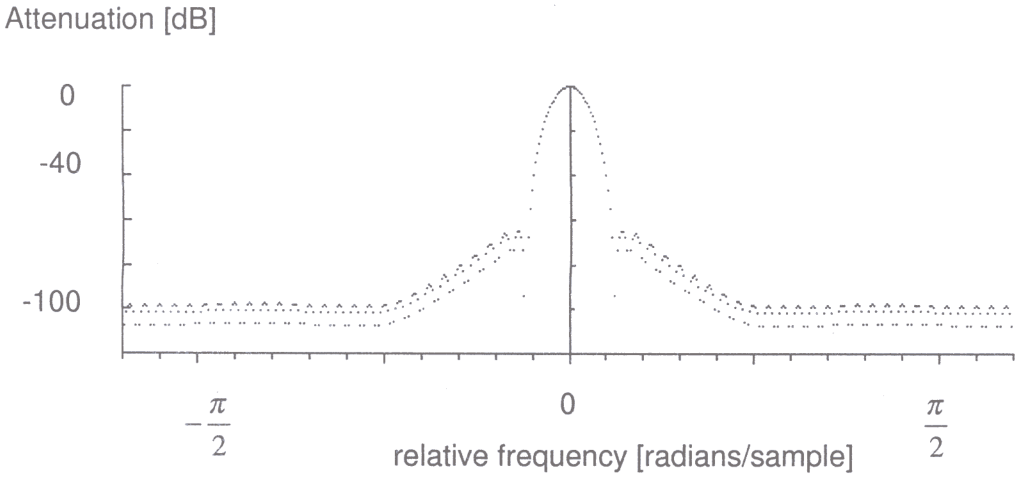

Fig. 2. Amplitude spectrum of the window used in the construction of the wavelet transform proposed.

The number of voices per octave (M in (9)) was chosen to be equal to 12, thus forming the 1/12 octave spacing of bands of frequency analysis, corresponding to the equally-tempered musical scale.

For computational reasons, the audio frequency range in the experimental system was reduced to from 30 Hz to 16 kHz. Also, the so-called scaling function, needed to represent the lower part of the frequency spectrum was not used, without any audible effects. This range encompasses 9 octaves. Together, the transform proposed analyses the signal in 108 (9 × 12) frequency bands.

The overlapping factor in the time domain, after initial experiments has been chosen at 75% for the highest voice in an octave. The wavelets for subsequent, lower voices are scaled to be longer by appropriate fractional powers of 2 and effective overlapping range increases. Four coefficients are computed for any signal frame equal to the length of the window.

The 1/12 octave *spacing* between the bands did not determine their *width*. According to the discussion in Sec. 7.1, the transform was designed to be redundant, which means not only overlapping of frame analysis in time, but also overlapping of bands in scale (frequency) domain. Thus, the appropriate balance between the resolutions in frequency and in time had to be found. The width of the bands was chosen so that the point of intersection of amplitude spectra of neighbouring bandpass filters was approximately at −3 dB. That width is determined by the length of the mother wavelet. The wavelet for the lowest scale used (shortest wavelet) was 48+1 samples long, with other wavelets in this highest octave being longer. The mother wavelet consisted of 16 windowed cycles of the complex sinusoid (real part). The experiments have been carried with the sampling frequency of 48 kHz, thus the centre frequency of the highest wavelet was equal to 16 kHz. The transform encompassed 9 octaves, and the length of the lowest scale wavelet in the bottom octave was 23196+1 samples and its frequency was centred at 33.1 Hz.

The reconstruction of the signal by inverse WT, for unit impulse test signal shows some distortion, resulting from the frame used still being not tight enough. The amplitude of these distortions can be estimated at below −40 dB in relation to the amplitude of the unit impulse. However, when tested on many different samples of audio signals, no audible difference has been heard by a group of listeners. The distortion could be reduced by increasing the redundancy, but then its rate would be impractically high. The WT presented above produces exactly 4 times more of data than there were samples in an input signal. In fact, this number can be reduced to 3 without audible effect by just less dense sampling of the time — scale plane. Higher reductions were obtained and informally tested using specific procedures relying on rules from psychoacoustics. These procedures could be a basis for algorithms for spectral redundancy. The experiments on such procedures are currently conducted.

Figure 3 shows the plot of the analysis of a 150 ms fragment of the recording of an orchestra. The result displayed is based on the modulus of complex coefficients obtained from the WT proposed. For better clarity, the plot presented is of the discrete, black and white type, instead of often used grey-scale type. The threshold used for classification black/white was gradually lowered along the frequency scale, to compensate for the usual decrease of energy of the acoustic signal towards higher frequencies. All frequency bands in the plot are double-pixel wide for better visualisation, thus all short vertical strips in the highest octave are showing single transform coefficients. Despite that the very dense musical fragment was deliberately chosen and slightly smearing nature of the black/white plot, concentration of energy in time-frequency plane is clearly visible, with some horizontal strips indicating strong harmonic components.

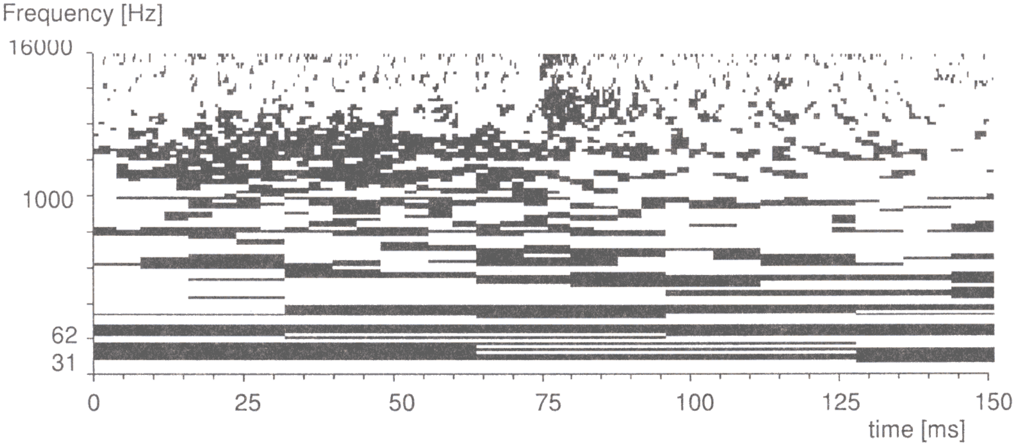

Fig. 3. The output of the transform proposed for a 150 ms fragment of orchestral music.

11. Conclusions

Analysis of the bibliography and the work performed by the author indicate, that successful modeling of the ear requires a specific linear frequency analysis procedure, which is a basis for any subsequent nonlinear operations. Its most important feature is constant Q bandpass filtering characteristics. One such method was proposed and tested. It relies on the Wavelet Transform with a modified Morlet mother wavelet, with oversampling the frequency domain by means of "voices". It performs well, but the data must be redundant by about four times, in order that the transform is perceptually invertible.

References

[1] A.N. Akansu, M.J.T. Smith [Eds.], *Subband and wavelet transforms*, Kluwer, Boston 1996.

[2] M. Bobrek, D.B. Koch, *Music segmentation using tree-structured filter banks*, Journal of the Audio Engineering Society, **46**, 5, 413–427 (1998).

[3] E. de Boer, *Classical and non-classical models of the cochlea*, J.A.S.A., **101**, 4, 2148–2150 (1997).

[4] J.C. Brown, *Calculation of a constant Q spectral transform*, J.A.S.A., **89**, 1, 425–434 (1990).

[5] C.S. Burrus, R.A. Gopinath, H. Guo, *Introduction to wavelets and wavelet transforms*, Prentice Hall, Upper Sadle River, NJ 1998.

[6] Y.T. Chan, *Wavelet basics*, Kluwer Academic Press, Norwell 1995.

[7] L. Cohen, *Time-frequency analysis*, Prentice Hall, Englewood Cliffs, NJ 1995.

[8] M.A. Cohen, S. Grossberg, L.L. Wyse, *A spectral network model of pitch perception*, J.A.S.A., **98**, 2, 862–885 (1995).

[9] A. Czyżewski, *New learning algorithms for the processing of old audio recordings*, 99th A.E.S Convention, Preprint no. 4078, New York 1995.

[10] I. Daubechies, *The wavelet transform, time-frequency localization and signal analysis*, IEEE Transactions on Information Theory, **36**, 5, 961–1005 (1990).

[11] Z. Engel, *Ochrona środowiska przed drganiami i hałasem*, Wydawnictwo Naukowe PWN, Warszawa 1993.

[12] J.L. Flanagan, *Models for approximating basilar membrane displacement*, The Bell System Technical Journal, 1163–1191 (1960).

[13] L.C. Gresham, L.M. Collins, *Analysis of the performance of a model-based optimal auditory signal processor*, J.A.S.A., **103**, 5, 2520–2529 (1998).

[14] F.J. Harris, *On the use of windows for harmonic analysis with the discrete Fourier transform*, Proceedings of the IEEE, **66**, 1, 51–83 (1978).

[15] W.M. Hartmann, *Pitch, periodicity, and auditory organization*, J.A.S.A., **100**, 6, 3491–3502 (1996).

[16] W. Heinbach, *Aurally adequate signal representation: The part-tone-time-pattern*, Acustica, **67**, 113–120 (1988).

[17] T. Irino, *A "Gammachirp" function as an optimal auditory filter with the Mellin transform*, Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 981–984, Atlanta 1996.

[18] T. Irino, H. Kawahara, *Signal reconstruction from modified auditory wavelet transform*, IEEE Trans. on Signal Processing, **41**, 12, 3549–3554 (1993).

[19] T. Irino. R.D. Patterson, *A time-domain, level dependent auditory filter: The gammachirp*, J.A.S.A., **101**, 1, 412–419 (1997).

[20] J.M. Kates, *A time-domain digital Cochlear model*, IEEE Trans. on Signal Processing, **39**, 12, 2573–2592 (1991).

[21] D.O. Kim, C.E. Molnar, R.R. Pfeiffer, *A system of nonlinear differential equations modeling basilar-membrane motion*, J.A.S.A., **54**, 6, 1517–1529 (1973).

[22] P. Kleczkowski, *Nowy sposób kodowania parametrów dla syntezy addytywnej sygnałów*, Mat. VII Sympozjum Inżynierii i Reżyserii Dźwięku, 125–128, Kraków 1997.

[23] J. Kovacevic, M. Vetterli, *Wavelets and subband coding*, Prentice Hall, Englewood Cliffs 1995.

[24] R.F. Lyon, *An analog electronic cochlea*, IEEE Trans. on A.S.S.P., **36**, 7, 1119–1133 (1988).

[25] H.S. Malvar, *Signal processing with lapped transforms*, Artech House, Boston 1992.

[26] R. J. McAulay, T.F. Quatieri, *Speech analysis/synthesis based on a sinusoidal representation*, IEEE Trans. on A.S.S.P., **34**, 4, 744–754 (1986).

[27] R. Meddis, M.J. Hewitt, *Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I. Pitch identification*, J.A.S.A., **89**, 6, 2866–2882 (1991).

[28] B.C. Moore [Ed.], *Frequency selectivity in hearing*, Academic Press, London 1986.

[29] S.H. Nawab, *Short-time Fourier transform*, [in:] J.S. Lim, A.V. Oppenheim [Eds.], Advanced Topics in Signal Processing, Prentice Hall, Englewood Cliffs, NJ 1988.

[30] B. Paillard, P. Mabilleau, S. Morisette, J. Soumagne, *PERCEVAL: Perceptual evaluation of the quality of audio signals*, Journal of the Audio Engineering Society, **40**, 1/2, 21–31 (1992).

[31] R.D. Patterson, *Auditory filter shapes derived with noise stimuli*, J.A.S.A., **59**, 3, 640–654 (1976).

[32] W.J. Pielemeier, G.H. Wakefield, *A high resolution time-frequency representation for musical instrument signals*, J.A.S.A., **99**, 4, 2382–2396 (1996).

[33] J.-C. Risset, D.L. Wessel, *Exploration of timbre by analysis and synthesis*, [in:] The psychology of Music, D. Deutsch [Ed.], Academic Press, New York 1982, 26–58.

[34] M.R. Schroeder, J.L. Hall, *Model for mechanical to neural transduction in the auditory receptor*, J.A.S.A., **55**, 5, 1055–1060 (1974).

[35] S. Shlien, *The modulated lapped transform, its time-varying forms, and its applications to audio coding standards*, IEEE Trans. on Speech and Audio Proc., **5**, 4, 359–366 (1997).

[36] M. Slaney, *Pattern playback in the '90s*, [in:] Advances in Neural Processing Systems 7, G. Tesauro, D. Touretzky, T. Leen [Eds.], Morgan Kaufmann Publishers, San Mateo, CA, 1995.

[37] T. Sporer, K. Brandenburg, *Constraints of filter banks used for perceptual measurement*, J.A.E.S., **43**, 3, 107–116 (1995).

[38] G. Strang, T. Nguyen, *Wavelets and filter banks*, Wellesley – Cambridge Press, New York 1996.

[39] R. Tadeusiewicz, *Sygnał mowy*, WKiŁ, Warszawa 1988.

[40] P.P. Vaidyanathan, *Multirate systems and filter banks*, Prentice Hall, Englewood Cliffs, NJ 1993.

[41] X. Yang, K. Wang, S. Shamma, *Auditory representations of acoustic signals*, IEEE Trans. on Information Theory, **38**, 2, 824–839 (1992).

[42] K. Wang, S. Shamma, *Self-normalizqtion and noise-robustness in early auditory representations*, IEEE Trans. on Speech and Audio Processing, **2**, 3, 421–435, July 1994.

[43] K. Wang, S. Shamma, *Spectral shape analysis in the central auditory system*, IEEE Trans. on Speech and Audio Processing, **3**, 5, 382–395 (1995).

[44] M.V. Wickerhauser, *Adapted wavelet analysis from theory to software*, IEEE Press, Piscataway, NJ 1994.

[45] E. Zwicker, H. Fastl, *Psychoacoustics, facts and models*, Springer-Verlag, Berlin 1990.

[46] E. Zwicker, T. Zwicker, *Audio engineering and psychoacoustics: Matching signals to the final receiver, the human auditory system*, J.A.E.S., **39**, 3, 115–125 (1991).