PERIODICA POLYTECHNICA SER. EL. ENG. VOL. 42, NO. 4, PP. 365–382 (1998)

THE CONTINUOUS WAVELET-TRANSFORM METHOD AND ITS APPLICATION TO SPEECH ENHANCEMENT

István PINTÉR

GAMF Technical College
Department of Informatics
6000 Kecskemét, Izsáki u. 10.
pinter@gandalf.gamf.hu

Received: January 5, 1998

Abstract

In this paper a new analysis method for nonstationary signals, the wavelet-transform is discussed. After a short introduction to continuous wavelet transform and to multiresolution analysis the concept of perceptual wavelets is introduced. Finally, possible application areas in digital speech processing are mentioned, followed by the experimental results of the perceptual wavelet-based speech enhancement.

Keywords: wavelet transform, speech representation, feature extraction, noise modelling, speech enhancement.

1. Introduction

With the emphasis on the method, this paper discusses the continuous wavelet transform (GOUPILLAUD et. al., 1984) through a special problem of digital speech processing (GORDOS – TAKÁCS, 1983).

Recently the wavelet transform is a well proven analysis method for nonstationary signals, and the algorithms derived from the wavelet theory became standards in digital signal processing (MEYER, 1993).

The universality of the method can be illustrated with successful applications from many kinds of scientific areas. Without the demand on completeness, the analysis of seismic signals (GOUPILLAUD et. al.,1984), the image processing (MALLAT, 1989, KISS et. al.,1994), the biomedical signal analysis for ECG (TUTEUR, 1990) and EEG (UNSER et. al.,1994), the analysis of $1/f$ noises (WORNELL, 1993), the vibration analysis in mechanical engineering (TANSEL et. al.,1993) and the speech analysis (KRONLAND-MARTINET et. al.,1987) can be enumerated as examples.

From the point of view of signal analysis methods the signals can be classified as stationary or nonstationary signals. When analysing former ones the Fourier-transform is a suitable method (namely decomposing the signals with complex exponential functions), and in the latter case the wavelet transform can be used (that is, deriving the signals f.e. as linear combinations of wavelets) (MEYER, 1993). The stationarity in the deterministic case can be defined with the time-independent instantenous amplitude and frequency (the instantenous amplitude is the amplitude of the so called complex analytical signal, and the instantenous frequency is the derivative of the phase of the latter), while the weakly stationary stochastic signals can be represented with time independent power density spectrum (MAMMONE, 1992). These properties are not valid for the non-stationary signals, thus a time-dependent description can only be given. Abrupt changes, signal transients can occur in this case, moreover their place (or time instant) cannot be predicted (FLANDRIN, 1990).

Recently there is a suitable method for non-stationary signals, the wavelet analysis. As important precedent on the one hand the Gábor's 'time-frequency atoms' (GÁBOR, 1946), on the other the orthonormal Haar-function system can be mentioned (HAAR, 1910). In the former case the speech can be decomposed into a sum of appropriate elementary signals, and in the wavelet-literature the Haar-system is the classical example of the so called dyadic wavelet base (DAUBECHIES, 1990).

Though – as a consequence of intensive research in this area – many wavelet functions have been found, nevertheless there is a question of how to choose the suitable wavelet for a specific signal processing problem. To solve this problem it is plauzible to begin with a wavelet, has been successfully applied to an analogous task, but the appropriate analysis wavelet function can be constructed starting from an adequate model of the physical system under investigation.

What follows in the rest of this article is a short introduction to continuous wavelet transform and multiresolution analysis in Section 2. After a review of application examples of wavelets in speech processing, Section 3 introduces the concept of perceptual wavelets, while Section 4 discusses the possible applications, among them the speech enhancement is details. Finally the summary and the acknowledgement can be found.

2. The continuous wavelet transform and the multiresolution analysis

When the problem is to analyse finer and finer details of the signals, the wavelet transform is an appropriate method. The analysis functions $w_{a,b}(t)$ can be derived by translation (shift) $b \in R$ and dilation $a > 0$ (the so called scale parameter) from the $w(t)$ wavelet with an energy-preserving transformation (COMBES, 1990):

$$w_{a,b}(t) = \frac{1}{\sqrt{a}} w\left(\frac{t-b}{a}\right) , \qquad a > 0 , \quad b \in R . \tag{1}$$

The $w(t)$ wavelet is well localised in time and in frequency, and the Fourier-transform $W(\omega)$ of $w(t)$ accomplishes the so called 'admissibility condition':

$$\int_{-\infty}^{0} \frac{|W(\omega)|^2}{\omega} \mathrm{d}\omega = \int_{0}^{+\infty} \frac{|W(\omega)|^2}{\omega} \mathrm{d}\omega = c < \infty , \qquad (2)$$

therefore $\int_{-\infty}^{+\infty} w_{a,b}(t) = 0$.

The special case of $a = 2^{-m}$; $b = n \cdot a$; $n, m \in Z$ corresponds to the discrete dyadic wavelets (Mallat, 1989):

$$w_n^m(t) = 2^{m/2} w\left(2^m t - n\right) . \qquad (3)$$

In the multiresolution analysis the functions $w_n^m(t)$ form an orthonormal basis.

Let's consider first the continuous wavelet transform which corresponds to wavelets in (1) and then the multiresolution analysis.

2.1. The continuous wavelet transform

The continuous wavelet transform can be defined by the integral below:

$$S(a, b) = \int_{-\infty}^{+\infty} s(t) \cdot w_{a,b}^*(t) \mathrm{d}t = \int_{-\infty}^{+\infty} s(t) \cdot w^* \left(\frac{t-b}{a}\right) \mathrm{d}t , \qquad (4)$$

where * denotes the complex conjugate.

The equation (4) can be interpreted in three ways. First, it can be considered as a scalar product of the signal $s(t)$ and the time-shifted version of the time-localised $w_{a,b}^*(t)$ analysing function, describing the signal details corresponding to scale a in $t = b$.

According to the second interpretation the signal $s(t)$ is analysed by a series of linear systems with impulse responses of the form $\frac{1}{\sqrt{a}} w\left(\frac{-t}{a}\right)$, so a wide variety of the descriptions of signal changes in $s(t)$ can be obtained from the slow ($a > 1$) to the rapid ($a < 1$) ones (the convolution integral interpretation of (4)).

As one can easily check, the wavelet transform $S(a, b)$ can be computed in the frequency domain with the inverse Fourier transform, too:

$$S(a, b) = \sqrt{a} \int_{-\infty}^{+\infty} S(\omega) W^*(a\omega) e^{jb\omega} \mathrm{d}\omega . \qquad (5)$$

It leads to a third interpretation because the argument of $W^*(a\omega)$ is in direct proportion to frequency at a given scale a. Thus taking the ratio of the bandwidth and the centre frequency, the ratio $\Delta\omega/\omega$ remains constant, so (4) essentially is a constant relative bandwidth (constant-Q) analysis.

In the case of sampled signals the computations can be accomplished with inverse DFT at different scales or with direct evaluation of a suitable approximation of (4):

$$S(n,a) = \frac{1}{\sqrt{a}} \sum_k s(k) \cdot w^* \frac{k-n}{a} \; ; \qquad k, n \in Z \; , \tag{6}$$

thus essentially by a linear convolution of $s(k)$ and $w^*(-l)$ (GORDOS – TAKÁCS, 1983; SIMONYI, 1984).

Having no other constraints for $w(k)$ the direct evaluation of (6) is very time-consuming because the length of the $w^*(-l)$ discrete filters is in direct proportion to the scale a. The calculations can be performed more quickly when the scale factor is the power of 2 (RIOUL – DUHAMEL, 1992), or in the case of the special B-spline wavelet with integer scales (UNSER et. al., 1994).

When processing bandlimited signals by choosing $a = a_0^k$, $k = 0, 1, \ldots, K-1$; $0 < a_0 < 2$, $|\omega| \in [\omega_1, \omega_2]$, the signal $s(t)$ can be analysed with K wavelets:

$$S\left(a_0^k, b\right) = S(k,b) = a_0^{k/2} \int\limits_\omega W^*\left(a_0^k \omega\right) \cdot S(\omega) \cdot e^{j\omega b} \mathrm{d}\omega \; . \tag{7}$$

In this case for the perfect reconstruction the condition

$$\sum_{k=0}^{K-1} a_0^{k/2} W^*\left(a_0^k \omega\right) = 1 \tag{8}$$

must be fullfilled:

$$s(b) = \sum_{k=0}^{K-1} S(k,b) = \sum_{k=0}^{K-1} a_0^{k/2} \int\limits_\omega W^*\left(a_0^k \omega\right) S(\omega) e^{j\omega b} \mathrm{d}\omega =$$

$$= \int\limits_\omega \left[\sum_{k=0}^{K-1} a_0^{k/2} W^*\left(a_0^k \omega\right) \right] \cdot S(\omega) e^{j\omega b} \mathrm{d}\omega \; . \tag{9}$$

The condition (8) plays important role when deriving perceptual wavelets.

Finally, several frequently used wavelet examples are given for practical applications.

The so called 'Grossmann-Morlet'-wavelet can be found in many publications covering different scientific problems (GOUPILLAUD et. al., 1984;

Tuteur, 1990; Kronland-Martinet et. al., 1987; Ambikairajah et. al., 1993):

$$w(t) = e^{\left(\frac{-t^2}{2} - j\omega_0 t\right)}; \qquad W(\omega) = e^{\frac{-(\omega-\omega_0)^2}{2}}; \qquad \omega_0 > 5.5\ . \tag{10}$$

This wavelet is interesting, because the Gábor's uncertainty-relation $\Delta\omega \cdot \Delta t \geq 0.5$ reaches equality in the case of functions in (10). The uncertainty Δx for the function $f(x)$ can be defined as (Gábor, 1946; Reid – Passin, 1992):

$$\Delta x = \frac{\int\limits_{-\infty}^{+\infty} (x-\overline{x})^2 \, |f(x)|^2 \mathrm{d}x}{\int\limits_{-\infty}^{+\infty} |f(x)|^2 \mathrm{d}x}\ ; \qquad \overline{x} = \frac{\int\limits_{-\infty}^{+\infty} x \cdot |f(x)| \mathrm{d}x}{\int\limits_{-\infty}^{+\infty} |f(x)|^2 \mathrm{d}x}\ . \tag{11}$$

The uncertainty $\Delta\omega$ can be computed from the Fourier-transform of $f(x)$ similarly.

Another possible example is the nth order B-spline wavelet (Unser et. al., 1994). In this case, the Fourier transform of the dilated version at scale m can be given by:

$$W_m(f) = m \cdot \mathrm{sinc}^{m+1}(m \cdot f)\ . \tag{12}$$

As it has been proven, an efficient, fast algorithm ($O(N)$ operations) can be given for this wavelet. Additional example is the so-called 'Mexican hat' function (Daubechies, 1990):

$$w(t) = \frac{2}{\sqrt{3}} \cdot \frac{1}{\sqrt[4]{\pi}} \left(1-t^2\right) \exp\left(\frac{-t^2}{2}\right)\ ; \qquad W(\omega) = \frac{2}{\sqrt{3}} \cdot \frac{1}{\sqrt[4]{\pi}} \cdot \omega^2 \cdot \exp\left(\frac{-\omega^2}{2}\right) \tag{13}$$

which has application f.e. in edge detection, but application examples of the

$$w(t) = e^{\frac{-|t|}{T}}\ ; \qquad W(\omega) = \frac{2T}{(\omega T)^2 + 1} \tag{14}$$

wavelet and of the classical Haar-wavelet

$$w(t) = \begin{cases} 1 & 0 \leq t < 0.5 \\ -1 & 0.5 \leq t < 1 \\ 0 & \text{else} \end{cases} \tag{15}$$

can be found in the related literature, too (Meyer, 1993).

In the latter case a 'multiscale analysis' has been elaborated as a dual pair of the multiresolution analysis (Stark, 1988). Because of its local and global resolution properties the Haar-system has been successfully applied in a 1D signal recognition problem earlier (Herendi, 1986; Pintér, 1986).

2.2. The multiresolution analysis

By means of multiresolution analysis (MEYER, 1993; WORNELL, 1993) the signal $s(t) \in V$ can be decomposed according to its changes by projections onto successive nested subspaces of the signal space V: $\dots V_m \subset V_{m+1} \subset \dots \subset V$. A particular subspace contains the signal details according to decomposition 2^m. The signal $s(t)$ is transformed from V onto V_m by the projection operator P_m, therefore the resulted signal is the 'best' approximation of $s(t)$. Because of the nested subspaces above the coarser details can be derived from finer ones.

The multiresolution analysis can be accomplished with a suitable $v(t)$ scale function; in this case for a given $m \in Z$ the functions

$$v_n^m(t) = 2^{m/2} v\left(2^m t - n\right) \ ; \qquad n \in Z \tag{16}$$

constitutes an orthonormal basis, so the approximation of $s(t)$ in this space is:

$$P_m\left[s(t)\right] = \sum_n a_n^m v_n^m(t) \ , \tag{17}$$

where

$$a_n^m = \int_{-\infty}^{-\infty} s(t) v_n^m(t) \mathrm{d}t \ . \tag{18}$$

The 'information loss' can be defined with the difference signal between two consequtive approximations:

$$D_m\left[s(t)\right] = P_{m+1}\left[s(t)\right] - P_m\left[s(t)\right] \ , \tag{19}$$

and $D_m[.] \in O_m$, the orthogonal complement of V_m in V_{m+1}: $V_{m+1} = V_m \oplus O_m$.

O_m is spanned by the

$$w_n^m(t) = 2^{m/2} w\left(2^m t - n\right) \tag{20}$$

orthonormal wavelet basis, therefore:

$$D_m\left[s(t)\right] = \sum_n b_n^m w_n^m(t) \ , \tag{21}$$

where

$$b_n^m = \int_{-\infty}^{+\infty} s(t) w_n^m(t) \mathrm{d}t \ . \tag{22}$$

When decomposing a particular signal to the scale 2^M, then by (17) and (19):

$$P_M\left[s(t)\right] = \sum_{m<M} D_m\left[s(t)\right] = \sum_{m<M} \sum_n b_n^m w_n^m(t) \ . \tag{23}$$

Equation (23) can be interpreted as an approximation does not contain the signal details finer, than 2^M. Thus in the case of $M \to \infty$ the signal $s(t)$ can be expressed as:

$$s(t) = \sum_m \sum_n b_n^m w_n^m(t) \ , \tag{24}$$

which is a decomposition of $s(t)$ according to a dyadic orthonormal wavelet basis.

The practical importance of the multiresolution analysis comes from the fact, that the coefficients a_n^m, b_n^m can be computed recursively. As a first step we need the value of a_n^{M+1}, which can be acquired by applying (17). After this downwards to other values m:

$$a_n^m = \sum_l h(l-2n) \cdot a_l^{m+1} \ , \tag{25}$$

$$b_n^m = \sum_l g(l-2n) \cdot a_l^{m+1} \ . \tag{26}$$

The corresponding reconstruction formula:

$$a_n^m = \sum_l [h(n-2l) \cdot a_l^m + g(n-2l) \cdot b_n^m] \ , \tag{27}$$

which gives a_n^M too, and thus $s(t)$ can be computed with a_n^m-s by applying (17).

The connection between the $h(n)$ and $g(n)$ sequences and the $v(t)$ scale function and $w(t)$ wavelet can be given by:

$$h(n) = \int_{-\infty}^{+\infty} v_n^1(t) \cdot v_0^0(t) \mathrm{d}t \ , \qquad g(n) = \int_{-\infty}^{+\infty} v_n^1 \cdot w_0^0(t) \mathrm{d}t \tag{28}$$

and

$$v(t) = \sqrt{2} \sum_l h(l) \cdot v(2t-l) \ , \qquad w(t) = \sqrt{2} \sum_l g(l) \cdot v(2t-l) \ ,$$

$$g(l) = (-1)^l h(1-l) \ . \tag{29}$$

In the case of the aforementioned, classical Haar-wavelet the values of $h(n)$ and $g(n)$ are: $h(0) = 1, h(1) = 1$; $g(0) = 1, g(1) = -1$. For practical applications there are many $h(n) - g(n)$ pairs (Mallat, 1989; Daubechies, 1990; Cody, 1992), moreover the multiresolution analysis can be accomplished with a VLSI chip (Cody, 1992).

3. Application of the wavelet transform to speech processing

When solving practical speech processing problems a widely-used speech model is the so-called quasi-stationary model (GORDOS – TAKÁCS, 1983). Accordingly, the speech is considered as a sequence of overlapping, quasi-stationary frames and the sampled speech is characterized by short-time parameters on the frame-by-frame basis. The short-time signal segment is fixed by a window-function in the time domain and because of the fixed window-length, the accessible frequency-resolution is limited, too (GORDOS – TAKÁCS, 1983; TARNÓCZY, 1984). Therefore, the localisation of speech transients can be achieved with limited accuracy only, as it has been demonstrated by several speech researchers (AMBIKAIRAJAH et. al.,1993). The localisation of signal transients, or abrupt changes is an important task in speech processing, because the (nearly) periodic opening/closing of the vocal chords during formation of a vowel is a similar event, so by the accurate event-localisation in time, the value of the fundamental frequency can be estimated or tracked more precisely. On the other hand the more adequate description of the nonstationary speech sounds is very important when desribing the fricatives, affricates, stops and when analysing the coarticulation process.

Comprehensibly, the interest of speech researchers has been aroused by the properties of wavelets, which has been strenghtened by the fact that the sound analysis mechanism of the inner ear can be modelled well with the constant-Q analysis – a particular property of the wavelet transform.

3.1. Wavelets and speech - an application overview

The selected speech processing applications are grouped below according to the applied wavelet-analysis technique.

The multiresolution analysis has been applied to determination of the fundamental frequency, in the presence of noise, too; the LeMarie-Meyer wavelet and the Mallat-algorithm has been used (KADAMBE – BOUDREAUX-BARTELS, 1992).

A general purpose speech analysis method has been elaborated on the theoretical basis of multiresolution analysis; the speech analysis is performed in the sequency-domain instead of frequency-domain (DRYJALGO, 1993). The method has been developed primarily for the analysis of speech transients; some of the published algorithms have been used earlier (HERENDI, 1986; PINTÉR, 1986).

In spite of the above mentioned success of the multiresolution analysis in speech processing, the so called speech-tailored wavelet analysis remains the subject of the further research. The fundamental reason can be found in the value of scale factor, which is in the case of multiresolution analysis exactly 2. Nevertheless, from the hearing theory the value of ≅0.8 would be expected (HERMES, 1993; SCHROEDER, 1993), so the continuous wavelet transform has become the subject of the research of speech-tailored wavelets.

Accordingly, the wavelet in (10) has been used for different purposes in speech research. The article with a new type of visual sound representation has become a classical one (KRONLAND-MARTINET et. al., 1987), and recently it has been reported, that in the transient-localisation problem this wavelet transform corresponds well to the analysis properties of the biophysical models of the inner ear (AMBIKAIRAJAH et. al., 1993).

For modelling the signal analysis properties of the human auditory system (TARNÓCZY, 1984) several different directions exist – with the corresponding wavelet constructions, of course. The continuous wavelet based functional model of the speech analysis properties of the basilar membrane in the inner ear has been proposed (IRINO - KAWAHARA, 1993); the analysing wavelet was derived from the measured transfer characteristic of the basilar membrane at a given location – the wavelet model corresponds well to the biophysical model of the basilar membrane. Moreover, there is a detailed. continuous wavelet-based model covering not only the basilar membrane-transformation, but the mechanical-nervous transduction process of (inner) hair cells and the cochlear nucleus signal processing as well (YANG et al.,1992); the wavelet function was derived from a transfer characteristic of the basilar membrane – inner hair cell system.

But there is another way to solve the speech-tailored wavelet analysis problem: namely the construction of special functions on the basis of the psychoacoustical properties of the hearing process. As an example, the FAM-functions can be mentioned, which are hearing-specific because of the applied frequency-warping function $g(x)$, characterizing the pitch-perception of a human listener (LAINE, 1992):

$$\mathrm{FAM}(n, g(x)) = \exp\left[\frac{1}{2}\ln\left(g'(x)\right) + j \cdot n \cdot g(x)\right] , \tag{30}$$

where $g'(x)$ is derivative of the $g(x)$.

3.2. The perceptual wavelets and their properties

The critical bands (ZWICKER, 1961; GREENWOOD,1961; TARNÓCZY,1984) play important role when constructing the perceptual wavelets. Two main interpretations of these bands exist. On the one hand, the ear sums up the energy in these frequency bands, other hand these are bands of equal length can be measured on the basilar membrane when investigating the tonotopic mapping of the latter. The place of these bands corresponds to the pure tone frequency nonlinearly, so this leads to the concept of nonlinear frequency-mapping or warping. It can be mentioned that – corresponding to the three different measurement method - there are three frequency warping functions: the Hz – Bark, Hz – ERB, Hz – mel; naming them with measurement units of the objective (physical) and subjective (perceptual) quantities.

Two of the available warping functions have been the most adequate for our purposes (PINTÉR, 1994a); the Traunmüller-formulae for Bark-scale (TRAUNMÜLLER, 1990) and the Greenwood's ones for the ERB-scale (GREENWOOD, 1961):

$$f^{[\mathrm{Bark}]} = 6.7\,\mathrm{asinh}\left(\frac{f^{[\mathrm{Hz}]}-20}{600}\right) \; ; \quad f^{[\mathrm{Hz}]} = 20 + 600\sinh\left(\frac{f^{[\mathrm{Bark}]}}{6.7}\right) \; , \quad (31)$$

$$f^{[\mathrm{ERB}]} = 16.7\lg\left(1+\frac{f^{[\mathrm{Hz}]}}{165.4}\right) \; ; \quad f^{[\mathrm{Hz}]} = 165.4\left(10^{0.06 f^{[\mathrm{ERB}]}}-1\right) \; . \quad (32)$$

With these in mind the basic idea in the construction of the perceptual wavelets is as follows: let's decompose the speech in the warped frequency scale with the help of functions of minimal uncertainty and unity-bandwidth, moreover the condition (9) must be met, too. (Or, from another point of view: let the signal analysis be optimal in Gábor-sense in the perceptual (subjective) scale, instead of the 'physical' (objective) frequency scale.)

Starting from the function $\exp(-c \cdot b^2)$ and defining the unity bandwidth between the 6 dB (50%) points, as one can easily check, the value of the parameter $c = 4\ln(2)$. Thus the analysing function at point b_0 – denoting the variable of the perceptual frequency scale with b –:

$$W(b) = \exp\left[-4\ln(2)\,(b-b_0)^2\right] = 2^{-4(b-b_0)^2} \; ; \qquad 0 < b_1 \leq b \leq b_2 \; , \quad (33)$$

where $b_0 = b_1 + k\Delta b$; $k = 0, 1, \ldots, K-1$; Δb is the distance between the consequtive maxima of the analysing functions when the condition (9) is met, and $[b_1, b_2]$ is the corresponding interval of the frequency band $[f_1, f_2]$ of the bandlimited signal $s(t)$ in the perceptual scale.

Returning to the Hz-scale two analysing wavelets can be found, according the two warping functions:

$$W_k^{\mathrm{Bark}}(f) = c_1 2^{-4[6.7\,\mathrm{asinh}[(f-20)/600]-(b_1+k\Delta b)]^2} \; , \quad (34)$$

where $b_1 = 6.7\,\mathrm{asinh}\left[(f_1-20)/600\right]$, and

$$W_k^{\mathrm{ERB}}(f) = c_1 2^{-4\left[7.253\ln\left(1+\frac{f}{165.4}\right)-(b_1+k\Delta b)\right]^2} \quad (35)$$

where $b_1 = 7.253\ln\left(1+\frac{f_1}{165.4}\right)$.

The values of c_1 and Δb have been found with the condition (9) numerically: $c_1 \cong 0.68008$, $\Delta b \cong 0.7526$ in both cases.

The analysing wavelets in the time domain can be computed with inverse Fourier-transform. The condition $W_k(-f)$ leads to complex analysing wavelet functions; when analysing the real speech signal with these wavelets two real output signal can be derived in each 'Bark-channel'. In order to derive real wavelets, the conditions $\mathrm{Im}\,[W_k(f)] = 0$; $W_k(f) = W_k(-f)$ or $\mathrm{Re}\,[W_k(f) =]\,0$; $W_k(f) = W_k(-f)$ must be met in the case of even or odd wavelet functions, respectively. When using real wavelets the speech signal $s(t)$ can be reconstructed from the the $s_k(t)$ decompositions with summation:

$$\sum_{k=0}^{K-1} s_k(t) = \sum_{k=0}^{K-1} \left[F^{-1}\left\{W_k(\omega)S(\omega)\right\}\right] = F^{-1}\left\{\sum_{k=0}^{K-1} W_k(\omega)S(\omega)\right\} =$$

$$= F^{-1}\left\{S(\omega)\sum_{k=0}^{K-1} W_k(\omega)\right\} = F^{-1}\left\{S(\omega)\right\} = s(t)\ . \qquad (36)$$

As it can be checked numerically, the scale property in (1) is approximated well in the case of Bark-wavelets and with high precision in the case of ERB-wavelets; the spectral condition of equation (2) is fulfilled by construction.

It is worth to note that when the construction is based on the nearly optimal function (REID - PASSIN, 1992), the $\cos^2(.)$, instead of the $\exp\left(-c \cdot b^2\right)$, the changes of the parameters c_1 and Δb are not important – this observation is useful when implementing the real-time version of the perceptual wavelet transform.

Referring to the above mentioned 'speech-tailored wavelet analysis' requirement, the construction above corresponds well to Scroeder's expectations (SCROEDER, 1993) concerning either the scale parameter or the value of the relative bandwidth. This latter can be defined as:

$$\frac{\Delta f}{f_k} = \frac{f\left(b_k + \frac{1}{2}\right) - f\left(b_k - \frac{1}{2}\right)}{\sqrt{f\left(b_k + \frac{1}{2}\right) \cdot f\left(b_k - \frac{1}{2}\right)}}\ ; \qquad (37)$$

where $f(b)$ denotes the inverse of the above mentioned frequency warping function. *Table 1* summarizes the expected and executed values.

Table 1. Expected values vs. those come from perceptual wavelets

	Schroeder	Bark-wavelet	ERB-wavelet
rel. bandwidth	0.15	0.15 ... 0.33	0.15 ... 0.22
scale factor (1/a)	1.15	1.11 ... 1.12	1.22

Some further interesting properties of the perceptual wavelets, the decomposition and time-localisation properties have been published elsewhere (PINTÉR, 1994b; 1996).

4. Application of perceptual wavelets to feature extraction and speech enhancement

4.1. Feature extraction and the visual representation of speech

As it has been shown in Section 3, the speech spectrum can be decomposed into a sum of K sub-spectrum – and the corresponding time-domain decompositions; the analysing wavelet can be characterized with critical bandwidth of unity and with a special shape in the frequency domain.

Because of the energy-summing properties of the critical bands it is plausible to describe the speech with the energy-levels in each perceptual-wavelet sub-band, respectively. These computations effectively result a feature vector, but better results have been achieved with the rms-values below:

$$e_k^m = \sqrt{\frac{\sum\limits_{n=0}^{N-1} \left[s_k^m(n)\right]^2}{N}}\ ; \qquad k = 0,\ 1,\ \dots,\ K-1\ ; \qquad m = 0,\ 1,\ \dots,\ M-1\ , \tag{38}$$

where e_k^m is the kth component of the feature vector in the mth N sample long speech frame. Because of the wavelet origin the computations can be accomplished on non-overlapping frames.

When describing the speech with these feature vectors in time, a new type of visual speech representation can be achieved, similar to the conventional spectrogram (but describing the nonstationary speech-details more accurately) and comparable to those published in the literature (DERMODY et. al., 1993; PINTÉR, 1996).

As it has been demonstrated with numerical experiments, these perceptual sound images are similar to those computed from the positive maxima of corresponding time-domain decompositions:

$$\varepsilon_k^m = \max_n \left[s_k^m(n)\right]\ ; \qquad k = 0,\ 1,\ \dots,\ K-1\ ; \qquad m = 0,\ 1,\ \dots,\ M-1\ . \tag{39}$$

This latter feature vector sequence corresponds to Mallat's wavelet-transform-maxima representation, therefore it can be considered as the basis of the further investigations concerning the speech compression problem.

Two examples of these feature vectors are given in *Fig. 1* and *Fig. 2* as illustrations.

4.2. A new speech enhancement method

It was shown in the previous section that the speech can be characterized well with the rms vectors in the perceptual sub-bands. In order to obtain a suitable noise model to the speech enhancement procedure it was interesting

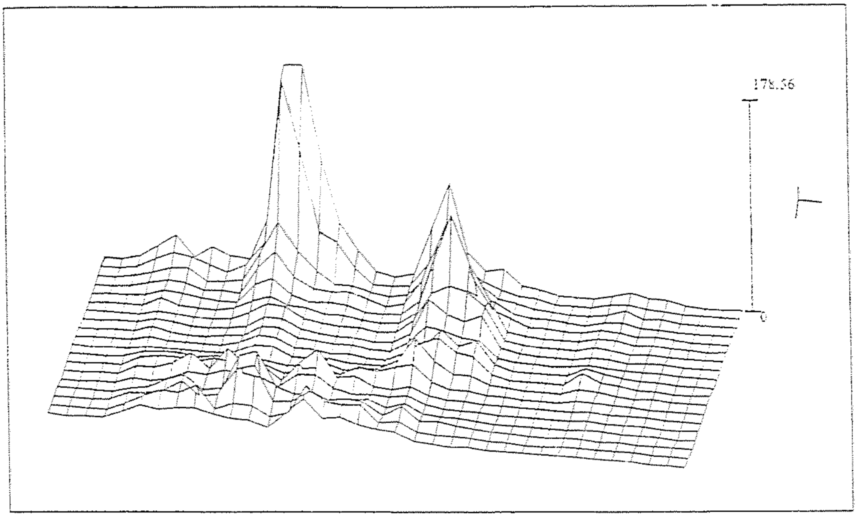

Fig. 1. Perceptual wavelet sub band rms-value representation of the Hungarian word 'sisak'

to analyse the noise signal with the same method. The numerical results of the analysis of six different noises show, that this description gives a good noise-discrimination, too, as it can be seen in *Fig. 3.*

Fig. 3 presents the average value $\overline{e}_k$ of the rms values according to (38), but in the enhancement process the variance σ_k is required, too. The results of *Fig. 3* are based on six types of bandlimited (300 ... 3400 Hz) noises. The appropriate noise and speech databases were built during the research and the latter are based on the written material of other speech processing problems (OLASZY, 1985; TAKÁCS, 1990).

The speech enhancement is based on the assumption that the noise can be characterized well with the estimated expected rms-value and its variance in each perceptual sub-band. During the noise suppression process only those sub-bands are involved into the reconstruction which exceeds the noise average. Further on, instead of this (implicit) step function the sigmoid-type sharpening function has been applied as nonlinearity:

$$T_k\left(e_k^m, \overline{e}_k, \sigma_k\right) = \frac{1}{1 + \exp\left\{-\frac{1}{2}\left[e_k^m - \left(\overline{e}_k + \frac{\sigma_k}{2}\right)\right]\right\}}, \qquad (40)$$

where e_k^m denotes the rms-value of the noisy signal in the kth sub-band of the mth speech frame. Thus the spectrum of the enhanced speech in the

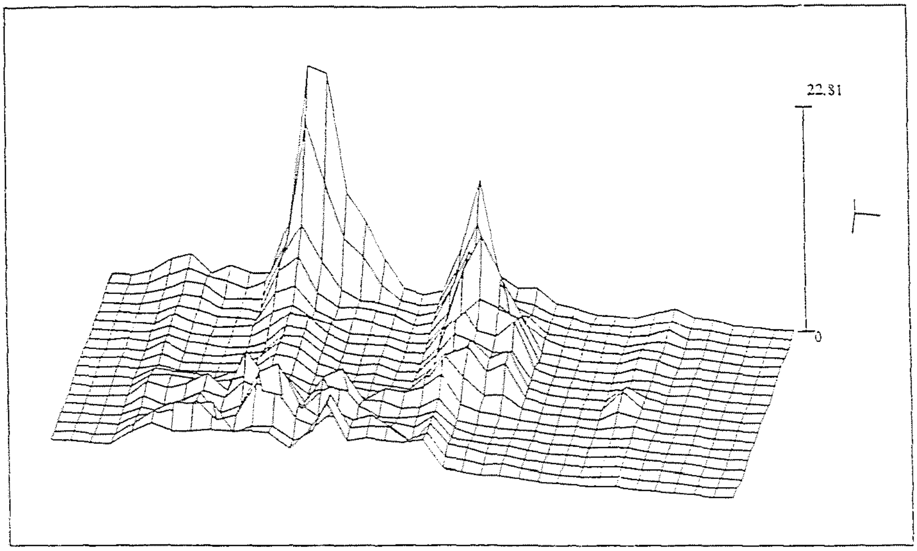

Fig. 2. Perceptual wavelet transform maxima representation of the Hungarian word 'sisak'

mth frame can be given by:

$$S_m(\omega) = \sum_{k=0}^{K-1} S(\omega) W_k(\omega) t_k \left(e_k^m, \bar{e}_k, \sigma_k\right) \ . \qquad (41)$$

The enhanced speech signal in the time domain can be computed with inverse Fourier transform in each non-overlapping speech frame, respectively.

In order to evaluate the performance of the speech enhancement method an objective measure of speech intelligibility would be necessary (GORDOS – TAKÁCS, 1983; TARNÓCZY, 1984; LEIJON, 1991; TARNÓCZY, 1995). The most adequate measure would be Tarnóczy's modified intelligibily index after a proper modification according to perceptual wavelet sub-bands; this would be the subject of another investigation.

Consequently, the performance of the speech enhancement method has been evaluated subjectively: the noise under question was added to the enhanced speech until it was perceived equally noisy as the original noisy speech. The improvement was then characterized with the segmental energy level of the subjectively added noise:

$$I = \frac{1}{M} \sum_{m=0}^{M-1} 10 \lg \sum_{n=0}^{N-1} s_m^2(n) \ . \qquad (42)$$

The experiments were carried on with six different noises and eight different noise amplitude in each case. The word 'bibe' has been selected from the

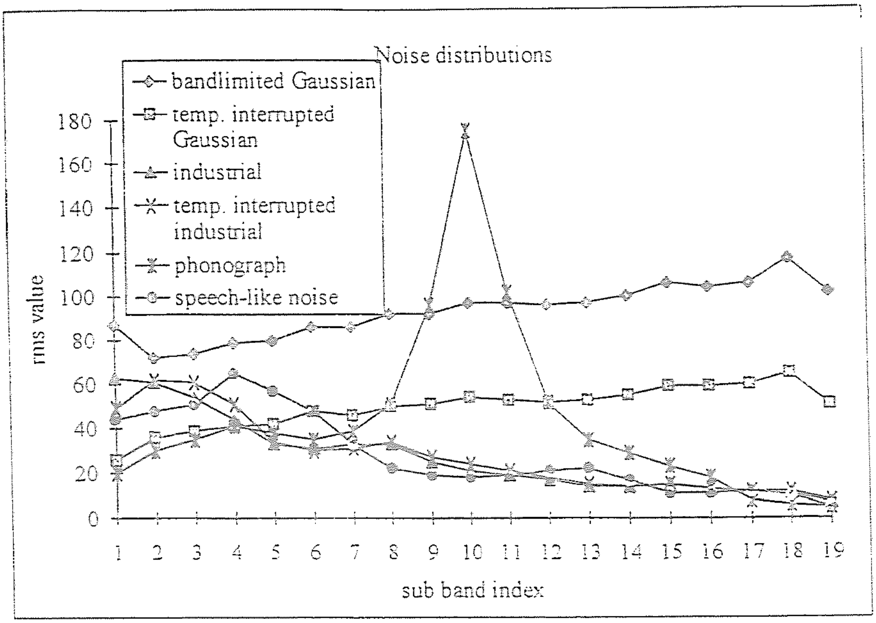

Fig. 3. Noise rms-distributions in percetual walvelet sub bands (300 ... 3400 Hz)

word database because its rms-distribution is similar to the most noise-rms distributions, therefore the performance limits of the enhancement procedure can be investigated.

Which is common in all cases is a saturation effect: above a noise-type-dependent noise level further improvements cannot be achieved: melodious artefacts are generated by the enhancement procedure. Describing the achievable improvement with this saturation limit value, the method resulted 26 dB improvement in the case of bandlimited Gaussian noise, 18 dB in the case of speech-like noise and 20 - 22 dB in all other cases. Therefore the improvement can be expected more than 18 dB, which means that these results correspond well to the published international results (PINTÉR, 1995).

5. Conclusions

In this paper the continuous wavelet transform, a new analysis method, suitable for nonstationary signals has been discussed. The possible interpretations and computation methods in time- and frequency domain has been presented too. For practical applications several wavelet functions were given, and the concept of perceptual wavelets has been introduced. Application examples for feature extraction, visual speech representation and noisy speech enhancement were given, too. In the latter case at least 18 dB improvement can be achieved in the case of six different noises.

It is planned to investigate the reconstruction of the speech from the above mentioned feature vectors, to realize the algorithms on a TMS320C30 DSP and to investigate an isolated speech recognizer in the case of noisy speech input. For designing the classifier part of the system several new results are available (Faragó et. al., 1993; Faragó - Lugosi, 1993; Delyon et. al., 1995).

Acknowledgements

This research was done in the Technical University of Budapest with the assistance of the Department of Telecommunication and Telematics. The author would like to thank to Prof. Géza Gordos for his kind help.

References

[1] Ambikairajah et. al., (1993): The application of the Wavelet Transform for Speech Processing. *Proc. Eurospeech*, pp. 151–154.

[2] Cody, M. A., (1992): The fast Wavelet Transform. *Doctor Dobb's Journal*, April pp. 16–28.

[3] Combes et al. eds, (1990): Wavelets. Springer Verlag.

[4] Daubechies, I., (1990): The wavelet transform, time-frequency localisation and signal analysis. *IEEE-IT*, Vol. 36. No. 5. Sept. pp. 961–1005.

[5] Delyon, B. et al., (1995): Accuracy Analysis for Wavelet Approximation. *IEEE Trans. on Neural Networks*, Vol. 6. No. 2. March pp. 332–348.

[6] Dermody, P. et al., (1993): Comparative Evaluations of Auditory Representations of Speech. Cooke, M., Beet, S., Crawford, M. eds.: Visual Representations of Speech Signals. John Wiley & Sons, pp. 229–236.

[7] Dryjalgo, A., (1993): Multiresolution Time-Sequency Speech Processing Based on Orthogonal Wavelet Packet Pulse Forms. *Proc. Eurospeech*, pp. 147–150.

[8] Faragó, A. - Lugosi, G., (1993): Strong Universal Consistency of Neural Network Classifiers. *IEEE-IT*, Vol. 39. No. 4. July pp. 1146–1150.

[9] Faragó, A. et al., (1993): Fast Nearest-Neighbor Search in Dissimilarity Spaces. *IEEE-PAMI*, Vol. 15. No. 9. Sept. pp. 957–962.

[10] Flandrin, P., (1990): Some Aspects of Nonstationary Signal Processing with Emphasis on Time-Frequency and Time-Scale Methods. in: Combes ed: Wavelets. Springer Verlag.

[11] Gábor, D., (1946): Acoustical Quanta and the Theory of Hearing. *Nature*, Vol. 169. pp. 591–602.

[12] Gordos, G. – Takács, Gy., (1983): Digital Speech Processing (in Hungarian) Műszaki Könyvkiadó, Budapest.

[13] Goupillaud, P. et al., (1984): Cycle-Octave and Related Transforms in Seismic Signal Analysis. *Geoexploration*, Vol. 23pp. 85–102.

[14] Greenwood, D. D., (1961): Critical Bandwidth and the Frequency Coordinates of the Basilar Membrane. *JASA*, Vol. 33. No. 10. Oct. pp. 1344–1356.

[15] Haar A., (1910): Zur Theorie der Orthogonalen Funktionensysteme. *Math. Annalen*, Vol. 69. pp. 331–371.

[16] Herendi, M., (1986): The Haar-Transformation and its Application. (in Hungarian) *Mérés és Automatika*, Vol. 34. No. 7. pp. 270–276.

[17] Hermes, D. J., (1993): Pitch analysis. in: Cooke, M., Beet, S., Crawford, M. eds.: *Visual Representations of Speech Signals.* John Wiley & Sons, pp. 3–25.

[18] Irino, T. – Kavahara, H., (1993): Signal Reconstruction from Modified Auditory Wavelet Transform. *IEEE-SP*, Vol. 41. No. 12. Dec. pp. 3549–3553.

[19] Kadambe, S. – Boudreaux-Bartels, G. F., (1992): Application of the Wavelet Transform for Pitch Detection of Speech Signals. *IEEE-IT*, Vol. 38. No. 2. March pp. 917-.924.

[20] Kiss, Cs. et al., (1994): Texture Analysis Based on Wavelet Decomposition. *Journal on Communications*, Vol. XLV. July-August pp. 47–50.

[21] Kronland-Martinet et al., (1987): Analysis of Sound Patterns Through Wavelet Transforms. *Intl. Journal of Pattern Recogn. and Artificial Intell.*, Vol. 1. No. 12. pp. 97–125.

[22] Laine, U. K., (1992): A new High Resolution Time-Bark Analysis Method for Speech. Preprint. Helsinki University of Techn. Acoustic Lab.

[23] Leijon, A., (1991): Optimization of Hearing-Aid Gain and Frequency Response for Cochlear Hearing Losses. *Technical Report*, No. 189. Chalmers University of Technology, Göteborg.

[24] Mallat, S. G., (1989): A theory for Multiresolution Signal Decomposition: the Wavelet Representation. *IEEE-PAMI*, Vol. 11. No. 7. July pp. 674–693.

[25] Mammone, R. J. ed., (1992): Computational Methods of Signal Recovery and Recognition. John Wiley & Sons.

[26] Meyer, Y., (1993): Wavelets: Algorithms and Applications. SIAM, Philadelphia.

[27] Olaszy, G., (1985): The Structure and the Synthesis of the most Frequent Elements of the Hungarian Speech. (in Hungarian) *Nyelvtudományi Értekezések*, 121. sz. Akadémiai Kiadó, Budapest.

[28] Pintér, I., (1986): On the Resognition of Signals from their Histograms' Haar-Spectra. (in Hungarian) *Gépgyártástechnológia*, Vol. 26. No. 8. pp. 362–368.

[29] Pintér, I., (1994a): Auditory Models in Speech Processing. (in Hungarian) *GAMF Közleményei*, Vol. XI. pp. 53–67.

[30] Pintér, I., (1994b): Perceptual Wavelet Representation of Speech Signals and the Enhancement of Noisy Speech. (in Hungarian) *Híradástechnika*, Vol. XLV. Sept. pp. 31–37.

[31] Pintér, I., (1995): Speech Enhancement by Soft Thresholding in the Perceptual Wavelet Domain. *Proc. IEEE Workshop on Nonlinear Signal and Image Processing*, Vol. II., pp. 666–669.

[32] Pintér, I., (1996): Perceptual Wavelet Representation of Speech Signals and its Application to Speech Enhancement. *Computer Speech and Language*, Vol. 10. No. 1. pp. 1–22.

[33] Reid, C. E. – Passin, T. B., (1992): Signal Processing in C. Addison-Wesley.

[34] Rioul, O. – Duhamel, P., (1992): Fast Algorithms for Discrete and Continuous Wavelet Transforms. *IEEE-IT*, Vol. 38. No. 2. March pp. 569–586.

[35] Schroeder, M. R., (1993): A Brief History of Synthetic Speech. *Speech Communication*, Vol. 13. No. 12. pp. 231–239.

[36] Simonyi, E., (1984): Digital Filters (in Hungarian) Müszaki Könyvkiadó, Budapest.

[37] Stark, H-G., (1988): Continuous Wavelet Transformation and Continuous Multiscale Analysis.*Preprint.* No. 146. Universit(t Kaiserslautern.

[38] Takács, Gy., (1990): Phonetic Recognition of Continuous Speech by Artificial Neural Network. (in Hungarian) Candidate Dissertation, Budapest.

[39] Tansel, I. N. et al., (1993): Monitoring Drill Conditions with Wavelet Based Encoding and Neural Networks. *Intl. J. Mech. Tools Manufact.* Vol. 33. No. 4. pp. 559–575.

[40] Tarnóczy, T., (1984): Sound Pressure, Loudness, Perceived Noise. (in Hungarian) Akadémiai Kiadó, Budapest.

[41] Tarnóczy, T., (1995): The Speech Intelligibility as Psyhoacoustical Concept. (in Hungarian) *Fizikai Szemle*, March pp. 90–97.

[42] TRAUNMÜLLER, H., (1990): Analytical Expressions for the Tonotopic Sensory Scale. *JASA,* Vol. 88. pp. 97–100.

[43] TUTEUR, F. B., (1990): Wavelet Transformation in Signal Detection. in: Combes et al. eds.: *Wavelets.* pp. 132–138. Springer-Verlag.

[44] UNSER, M. et al., (1994): Fast Implementation of the Continuous Wavelet Transform with Integer Scales. *IEEE-SP,* Vol. 42. No. 12. Dec. pp. 3519–3523.

[45] WORNELL, G. W., (1993): Wavelet-Based Representations for the 1/f Family of Fractal Processes. *Proceedings of the IEEE,* Vol. 81. No. 10. October pp. 1428–1450.

[46] YANG., X. et al., (1992): Auditory Representation of Acoustic Signals. *IEEE-IT,* Vol. 38. No. 2. March pp. 824–839.

[47] ZWICKER, E., (1961): Subdivision of the Audible Frequency Range into Critical Bands (frequenzgruppen). *JASA,* Vol. 33. No. 2. Feb. p. 248.