WARPED LINEAR PREDICTION (WLP) IN SPEECH AND AUDIO PROCESSING

Unto K. Laine, Matti Karjalainen, and *Toomas Altosaar*

Helsinki University of Technology, Acoustics Laboratory
Otakaari 5A, SF-02150 Espoo, Finland

ABSTRACT

In this paper linear prediction process is applied to frequency warped signals. The warping is realized by using orthonormal FAM[1] functions. The general formulation of WLP is given and effective realizations with allpass filters are studied. The application of auditory WLP to speech coding and speech recognition has given good results.

1. INTRODUCTION

In the field of speech and audio processing many proposals have been made to compress the signals according to the human auditory system. The earliest method of speech coding was the use of a filterbank with gradually increasing channel bandwidths [1]. Typically a total compression of ten to one or even more can be achieved with this method [2, 3]. Many speech recognizers have adopted the same principle of a nonuniform resolution preprocessing, e.g., [4]. Perfect reconstruction orthonormal filterbanks and wavelet-based techniques have been recently most actively studied [5].

Nonuniform resolution FFT was introduced by Oppenheim, Johnson and Steiglitz [6, 7]. The main idea of their study was to use a network of cascaded first order allpass sections for frequency warping of the signal and then to apply standard FFT to produce the warped spectra from the preprocessed signal. Very similar structures were used by Lee [8] to produce orthonormal bases like Laguerre, Fourier, and Legendre for frequency domain filter design.

The idea of the warped FFT was then applied to warped linear prediction (WLP) by Strube [9]. In his study a cascaded first order allpass network produces the frequency warping corresponding to the auditory Bark scale. The autocorrelation function for LP is calculated from the warped signal. The warping leads to an LP process which gives directly the auditory scale representation for the signals. This method was later applied to an ADPCM system [10].

In this paper a new formulation of the WLP is made based on the class of FAM functions. It is shown that the use of allpass sections leads to very efficient computation and that the warping can as well be realized by second or even higher order allpass sections.

2. FAM CLASS OF ORTHONORMAL FUNCTIONS

The fam function $\phi_\nu(x,a)$ is defined by (1), where $j = \sqrt{(-1)}$ and $\nu'(x)$ denotes the first derivative of the function $\nu(x)$ [11, 12].

$$\phi_\nu(x,a) = \sqrt{\nu'(x)}\, e^{j 2\pi a \nu(x)} \qquad (1)$$

[1] FAM denotes the class of Frequency-Amplitude Modulated complex exponentials. The name *fam* is used like *sine* and *cosine* when referring to an individual member in the *FAM class.*

In the continuous (non-discrete) case $x, \nu \in \mathbf{R}$ and $a \in \mathbf{R}$ or $\mathbf{Z}$. In the following we assume that the a-domain is discrete ($a \in \mathbf{Z}$). The variable a can be associated to the order of the function. Generally (1) is also valid when x- and ν-domains are discrete.

The inner function $\nu(x)$ can have many interpretations. It can be called a *generative* function because it specifies the central properties of the actual set. If we use fam functions as a basis in a linear transform then $\nu(x)$ defines a new *scale* on which the information is mapped. Typically the mapping is *nonlinear* causing a scale *warping* and also a change in the *resolution* of the description. Therefore $\nu(x)$ could also be called a warping function, or resolution function.

Fam function exists only when $\nu(x)$ fulfills certain conditions. Generally speaking it must be well behaved and its first derivative must exist almost everywere. In many practical applications, e.g. in signal processing, $\nu(x)$ must be square-integrable (in Lebesgue sense). In (1) it is assumed that $\nu(x)$ is monotonically increasing (having a positive derivative) over the range of orthogonality. A more general formula is easily achieved by taking the absolute value of the derivative of $\nu(x)$. Then, functions with nonmonotonic $\nu(x)$ can also be used to produce orthonormal sets. Note that when $\nu(x) = x$, (1) reduces to the Fourier kernel.

2.1. Orthonormality

The fam functions are orthonormal when (2) is valid

$$\int_{x_1}^{x_2} \phi_\nu(x,a)\,\phi_\nu^*(x,b)\,dx \begin{cases} =1 & b=a \\ =0 & b\neq a \end{cases} \qquad (2)$$

The range of orthogonality over x is assumed to be $[x_1, x_2]$. By noting that $\nu'(x)=d\nu/dx$ we can change the integration variable and show that by proper choice of $\nu'(x)$ (2) is valid [12]. In the following $\phi(x,a)$ functions are mainly used to warp the frequency domain ($x = f$). $\nu(x)$ is typically normalized according to Fig. 1.

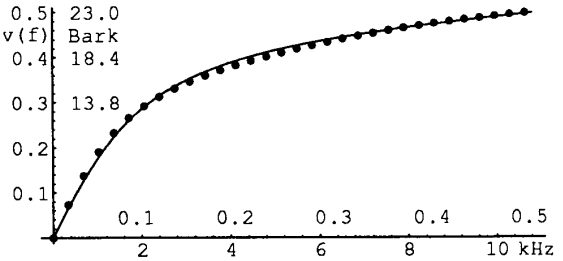

Fig. 1. Normalized Hz-to-Bark warping function $\nu(x)$ (dots) approximated by a first order allpass filter ($\lambda = 0.62$, 0-11 kHz).

2.2. Famlet Class of Orthonormal Functions

Fam functions have up to now found interesting applications in signal analysis and representation as well as in filter design [14].

0-7803-1775-0/94 $3.00

In all these cases the functions are defined in the frequency domain. Sometimes it is practical to work with their representations in the time domain, too. These functions which are called *famlets* also form an orthonormal class. The name famlet was chosen to describe them as a type of FAM based, time domain, compact waveforms having partially similar use as wavelets. Famlets are defined by:

$$\psi_\nu(t,a) = \mathbf{F}^{-1}\,\phi_\nu(f,a) \qquad (3)$$

In (3) time domain famlets are produced simply by inverse Fourier transforming the frequency domain fam functions.

2.3. Class of FAM and Famlet Transforms

Like many known orthonormal bases FAM basis can also be used to produce integral transforms. For our present purpose the FAM transform is formulated in the following way:

$$\hat{\Phi}_\nu S(f) = \int_{f_1}^{f_2} S(f)\,\phi_\nu^*(f,a)\,df = s_\nu(a) \qquad (4)$$

The range of orthogonality is $[f_1, f_2]$. Note that when $\nu(f) = f$ the FAM transform reduces to the Fourier transform. The Fourier case is one of the infinitely many FAM transforms and the only one where no warping occurs. Correspondingly we can define the FAMlet transform.

$$\hat{\Psi}_\nu s(t) = \int_{t_1}^{t_2} s(t)\,\psi_\nu(t,a)\,dt = s_\nu(a) \qquad (5)$$

Note that the famlet transform of the signal $s(t)$ is equal to the FAM transform of its spectrum $S(f)$. The new signal $s_\nu(a)$ represented in the a-domain (transform domain) is the frequency warped version of the original signal $s(t)$.

2.4. Continuous vs. Discrete Cases

Up to this point we have discussed continuous functions, integral transforms, and related operators. In the following we change the notation slightly in order to work with the corresponding discrete versions. This means, e.g., that integral operators are changed to matrix operators. Also, signals and their spectra are now represented in the indexed, discrete domains.

$$\begin{aligned} &\hat{\Phi}_\nu \rightarrow \Phi_\nu \qquad s(t) \rightarrow s(n) \qquad t, f \in \mathbf{R} \\ &\hat{\Psi}_\nu \rightarrow \Psi_\nu \qquad S(f) \rightarrow S(k) \qquad n, k \in \mathbf{Z} \end{aligned} \qquad (6)$$

2.5. Completeness and Unitarity

One fundamental difference between the continuous and discrete cases must be kept in the mind. When x-domain is continuous from $-\infty$ to $+\infty$ the FAM sets are complete [12]. Any function in this domain can be represented with increasingly high accuracy with the fam functions just by adding new terms in the expansion. This is not the case when we deal with limited ranges in discrete domains, e.g., $s(n)$ $n \in [0, N\text{-}1]$ or $S(k)$ $k \in [-K, \ldots, 0, \ldots, K\text{-}1]$. In these cases *the FAM sets are not complete*, or, the range of completeness is only a fraction of the range of orthogonality. We may also say that the completeness is *frequency dependent*. There is *only one complete discrete set* of fam functions and it is the *discrete Fourier basis*. All other nonlinear or warped bases are not complete. This follows from the fact that warping makes the derivative of the ν function larger than 1 in some frequency range (see Fig. 1) which leads to earlier folding of the discrete ϕ functions when their order (a) increases. *The discrete fam set - when defined over a limited range - will always include less members than the corresponding discrete (uniform and linear) Fourier basis.* This also means that the transform matrices are not square matrices and the transformation leads to *information compression*. Original information given on a uniform scale is compressed when mapped on a nonuniform scale and the new representation is not a complete description of the original one. In auditory types of processing the compression leads to frequency resolution reduction at higher frequencies (Fig. 1).

It must also be noted that in the discrete case when the FAM transform is defined over a limited range *it is not a unitary operator* as is the case when an infinite range is used. Unitary operators have two important features: they preserve inner products, that is $\langle Ug, Uh\rangle = \langle g, h\rangle$, where $g, h \in l^2$, the Hilbert space of square summable functions, and U denotes a unitary operator. Secondly, they map orthonormal bases to orthonormal bases. These properties guarantee that *Parseval's theorem* is valid under these operators. More clearly: the energy of a signal measured on any of these domains equals $(\langle s(n),s(n)\rangle = \langle S(k),S(k)\rangle = \langle s(a),s(a)\rangle)$. When discrete, limited range FAM or famlet transforms are used *Parseval's theorem is not valid*. However, our numerical simulations show that the inner product is approximately preserved when the transform domain is scaled according to the number of fam functions used and when the weight $(d\nu/dx)$ is ignored.

3. WARPED SPECTRUM AND WARPED SIGNAL

FAM and famlet transforms can be used to produce new spectral representations with variable (frequency dependent) spectral resolution [13]. These transforms have been utilized, e.g., to produce auditory spectra and spectrograms in which the linear frequency scale is mapped (warped) to a psychoacoustic frequency scale like Bark or ERB-rate (Fig. 1). The method is shortly as follows: first define the warping function from Hz to Bark (or ERB-rate), i.e., $\nu_j = \nu(f_k) = \nu(k)$, where f_k denotes a discrete frequency point and k the corresponding index in the linear frequency scale (Hz). Then construct the set (1) of orthonormal fam functions. This set of functions produces the FAM transform matrix Φ_ν, which is then used to map the Fourier spectrum $S(k)$ of the signal $s(n)$ to the frequency warped signal $s(a)$. Finally, use the Fourier transform to produce the auditory spectrum (spectrum on the new, warped ν-scale). This procedure is formally given by (7), where $\mathbf{F}$ denotes the Fourier transform and Φ the FAM transform.

$$S(\nu_j) = \mathbf{F}\,\Phi_\nu S(k) = \mathbf{F}\,s_\nu(a) \quad j,k,a \in \mathbf{Z} \qquad (7)$$

According to (5) the warped signal $s(a)$ can be also produced by the famlet transform which maps the original signal $s(n)$ to the a-domain $s(a)$. When this is Fourier transformed as above we get the same warped spectrum $S(\nu)$ as in (7). When the order of the famlet and Fourier transforms is changed and the Fourier operator works on the famlet transform matrix instead of the warped signal an orthogonal filterbank is produced which has a nonuniform frequency resolution according to the warping $\nu(f_k)$ [13].

4. FAM TRANSFORM BY ALLPASS FILTERS

Up to now we have described a general theory and methodology for frequency scale warping with fam functions. Almost any type of warping can be treated within this general framework. However, from the implementation point of view there is presently no means to realize these processes in a highly efficient way. Many arithmetic operations in the form of FFT and FIR computation are needed. It is easy to see that if we limit the warping functions to those cases where the complex exponential part of fam can be realized by *allpass filters*, it is possible to replace the expensive FIR type of processing by a more efficient recursive IIR process.

4.1. Warping with First Order Allpass Sections

The fam functions defined in the frequency domain are now presented in the form

$$\phi_\nu(f,a)=\sqrt{\nu'(f)}\,e^{j2\pi a\,\nu(f)}=\sqrt{M(z)}\,\{\Theta(z)\}^a \qquad (8)$$

where M(z) denotes the magnitude weighting function for the allpass filters. M(z) is assumed to have zero phase. This is not a necessity because the fam set (8) is orthonormal even if sqrt[M(z)] has a nonzero phase, i.e., the set can be multiplied by some phase warping function without affecting its orthonormality. The complex exponential of fam is now modeled by an allpass filter $\Theta(z)$.

By starting from the orthogonalization of the impulse responses of cascaded identical first order allpass sections the frequency domain weighting function M(z) can be derived. The process leads to fam functions of the form

$$\begin{aligned}\phi_{\nu_1}(z,a)&=\sqrt{M_1(z)}\,[\Theta_1(z)]^a \qquad \lambda<1\\ &=\sqrt{\frac{(1-\lambda^2)z^{-1}}{(1-\lambda z^{-1})(z^{-1}-\lambda)}}\left[\frac{z^{-1}-\lambda}{1-\lambda z^{-1}}\right]^a\end{aligned} \qquad (9)$$

In (9) M(z) is purely zero phase because it has symmetrical poles with respect to the unit circle: one pole in the unit circle ($\lambda < 1$) and one symmetrically outside the unit circle. Additionally it has a phase compensating pole at the origin which finally changes this filter from linear phase to zero phase. Now, by solving the phase characteristics of the first order allpass filter and comparing its derivative to $M_1(z)$ it can be shown that (9) equals the fam functions in (1), i.e., $\nu'(f) = M_1(z)$.

The pole locations of $M_1(z)$ mean that the function is unstable. However, according to the fam theory we can now introduce a phase to the weighting function by moving the pole inside the unit circle and neglecting the pole at the origin. This leads to a new orthonormal basis (10), which corresponds to the classical discrete version of z-transformed Laguerre functions.

$$\tilde{\phi}_{\nu_1}(z,a)=\frac{\sqrt{(1-\lambda^2)}}{(1-\lambda z^{-1})}\left[\frac{z^{-1}-\lambda}{1-\lambda z^{-1}}\right]^a \qquad (10)$$

We have shown that unitary warping of spectra can be performed by using cascaded first order allpass filters to form an orthonormal basis in the frequency domain. The warped signal appears at the output taps of this allpass chain. The signal can be warped and $S(k)$ formed by using this orthonormal base. If we use M(z) as the weight then $S(k)$ is produced directly with the allpass sections without any weighting function.

4.2. Warping with Second Order Allpass Sections

The use of first order allpass sections allows the largest warping to take place only close to the DC point or at the folding frequency. If the point is desired to be located elsewhere on the frequency axis we have to replace λ with a complex number and compute with complex valued signals. Another possibility is to use second order allpass sections. These filters produce an average delay of two samples which leads to folding. This can be avoided by computing the chain with two times higher sampling frequency and by inserting zeroes between the samples of the input sequence. Another way is to choose a new filter structure where one delay element is compensated by a delay in the parallel branch. This is the case in the Saramäki-Renfors filter [15].

One aspect in the WLP is the realization of the synthesis filter. Allpass sections are delayless and can not be used as such in the recursive synthesis structure [9]. However, according to Steigliz [16] both the first and the second order allpass sections can be modified to form a stable, recursive, synthesis filter.

5. WARPED LINEAR PREDICTION

Linear prediction (LP) is a process which, based on the statistics of the signal $s(n)$ to be predicted, gives optimized coefficients (vector α) for a FIR type predictor $P(z)$ of order p [17]. The optimization is based on the minimization of the average squared prediction error (squared difference between the actual and the predicted value). When the predictor knows p samples from the history of the signal it produces a predicted value for the coming, new sample. This is made as a *linear combination* of p past samples weighted by the predictor tap coefficients. The LP process can be formulated in many different ways. Autocorrelation, covariance, and lattice formulations are the most commonly used.

Autocorrelation method of the LP can be seen as a process or nonlinear operator which produces predictor coefficients from the autocorrelation function $R(m)$ of the signal $s(n)$. The warped linear prediction (WLP) is now defined by:

$$\begin{aligned}&\alpha_\nu = L_{ac}\,R_{s_\nu}(b) \qquad b\in a\in Z\\ &\alpha_\nu=(\alpha_{\nu_1},\alpha_{\nu_2},\ldots,\alpha_{\nu_p})^T\end{aligned} \qquad (11)$$

Note that the warped autocorrelation function (WAC) and the correlation lag b are now defined in the a-domain. The main problem which remains is: how to produce the WAC.

6. WARPED AUTOCORRELATION

The set of equations (12) summarizes different methods to compute the warped autocorrelation function (WAC).

$$\begin{aligned}&(i)\quad R_{s_\nu}(b)=\sum_a s_\nu(a)\,s_\nu(a+b)\\ &(ii)\quad R_{s_\nu}(b)=\Phi_\nu[P(k)]\\ &(iii)\quad R_{s_\nu}(b)=\Psi_\nu[R(m)]\\ &(iv)\quad R_{s_\nu}(b)=\sum_n s(n)\,s_\nu(b,n)\end{aligned} \qquad (12)$$

The first "direct form" is computationally expensive because the whole frame of the warped signal has to be derived first. In the second method the power spectrum of the unwarped signal is first computed and then FAM transformed to WAC. This method is clearly more efficient than the first one. However, an FFT and an absolute value squared process is needed as well as a real valued FAM transform. The third method has about the same complexity as the second one. Now the conventional autocorrelation function (AC) is first derived and then famlet transformed. Because the famlets are relatively long functions of time we need several times more points in the AC than there are in the WAC (see also Strube [9]).

A new method is given by (iv). It leads to a running computation where $s(n) = s(0,n)$ is the unwarped signal at the input of the allpass chain and $s(b,n)$ the warped component picked up at the bth output tap of the chain. This method leads to very efficient recursive computation and it can be proven that when a frame of limited length is processed this WAC is exactly the same as given by the method (ii). The recursion in the allpass sections does not make any harm because the length of the sequence $s(n)$ is limited by the analysis frame.

WLP utilizes auditory spectra in a similar way with PLP by Hermansky [4]. However, one notable difference is that in PLP the loudness scaling of the power spectrum can be handled whereas in WLP only linear transforms are used.

7. APPLICATIONS OF WLP

One of the most natural applications of frequency domain warping is the modeling of the auditory system [19]. Here we show how WLP can succesfully and efficiently be applied to Bark-scaled representation and coding of speech and audio signals.

The Bark scale is a commonly used mapping to describe the frequency resolution of the human auditory system. Less resolution is needed for high frequencies because the critical bandwidth for resolving signal components is broader for higher frequencies. Fig. 1 shows the Bark curve and its approximation by first order allpass warping (λ=0.62) for a signal bandwidth of 11 kHz. A remarkable reduction in linear prediction filter order can be achieved theoretically by WLP over normal LP: approximately 8 to 1 for full audio bandwidth, about 3.4 to 1 for a 7 kHz bandwidth, and about 2.1 to 1 for a speech bandwidth of 3.4 kHz.

We have carried out preliminary experiments in using WLP for audio coding with multipulse excitation and also studied the applicability of the method to wideband speech coding. Fig. 2 shows the spectral flatness of the prediction error signal of WLP and normal LP as a function of filter order for vowel /e/ measured over the frequency range 0-3 kHz [17]. In this case the WLP of order 9 gives the same flatness as the conventional LP of order 20. The coding of the residual or excitation signal remains the same as in normal LP.

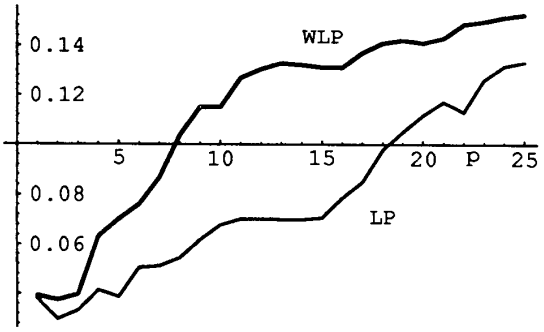

Fig. 2. Spectral flatness in the range 0-3 kHz of the residual signal as a function of filter order in normal and warped linear prediction of vowel /e/.

Another application experiment was to use WLP as a preprocessor to speech recognition. Running warped autocorrelation (formula 12 *iv*) with a 25 ms Hamming prewindowing was computed by the first-order allpass method (λ=0.62, sampling rate of 22.05 kHz) and the result of WLP was represented by lattice filter coefficients (reflection coefficients). A set of such coefficient vectors collected over a time frame was used as the input to a neural network classifier.

A multi-layer perceptron neural net was trained by back-propagation to act as a diphone detector to find and coarse classify Finnish stop-vowel diphones [18]. In a reference system we used the output of our standard preprocessor, an auditory spectrum vector [19] of 24 elements (24 Barks). The size of the Bark-warped WLP vector was 9 elements. A time frame (like in Time-Delay Neural Nets) of 7 vectors was collected over a time interval of 120 ms in both cases so that the network input dimensions were 7 x 24 for auditory spectrum and 7 x 9 for WLP coefficients. The neural nets had one hidden layer of 4 to 20 hidden nodes; this number was not found critical. The output of the net was used to detect the position of stop-vowel diphones in isolated word speech signals. The training material consisted of about 5000 diphones and a test set containing 620 diphones including 114 stop-vowel units was used in the experiment.

The reference system (auditory spectrum input) resulted in a 1.3 to 2.2 % error rate (average 1.7 %) and the corresponding error rate for the WLP input was 1.6 to 2.1 % (average 1.9 %). (Error rate = percentage of deleted and inserted stop-vowels). As a conclusion the WLP preprocessor performed almost as welll as the auditory spectrum in spite of a radical reduction (2.6 to 1) in data size. The computation of WLP is also several times faster that that of the auditory spectrum.

ACKNOWLEDGMENT

This study has been financed by the Academy of Finland.

REFERENCES

[1] Dudley H., "The vocoder", Bell Labs Record, 17, pp. 122-126, 1939.

[2] Flanagan J. L., *Speech Analysis, Synthesis and Perception*, Second Edition, Chapter 8, pp. 321-385, Springer-Verlag, New York, 1972.

[3] Gold B., Rader C. M., "The channel vocoder", IEEE Trans. Audio and Electroacoustics, AU-15, No. 4, pp. 148-160, Dec. 1967.

[4] Hermansky H., "Perceptual linear predictive (PLP) analysis of speech", J. A. Soc. Am., 87 (4), pp. 1738-1752, April 1990.

[5] Akansu A. N., Haddad R. A., *Multiresolution Signal Decomposition*, Academic Press Inc., Boston, 1992.

[6] Oppenheim A. V., Johnson D. H., Steiglitz K., "Computation of spectra with unequal resolution using the fast Fourier fransform", Proc. of the IEEE, 59, pp. 299-301, Feb. 1971.

[7] Oppenheim A. V., Johnson D. H., "Discrete representation of signals", Proc. of IEEE, 60, No. 6, pp. 681-691, June 1972.

[8 Lee Y. W., *Statistical Theory of Communication*, ch. 19: "Synthesis of optimum linear systems by means of orthonormal functions", Wiley, New York, 1960.

[9] Strube H. W., "Linear prediction on a warped frequency scale", J. Acoust. Soc. Am., 64 (4), pp. 1071-1076, Oct. 1980.

[10] Krüger E., Strube H. W., "Linear prediction on a warped frequency scale", IEEE Tr. on Acoustics, Speech, and Signal Processing, 36 (9), pp. 1529-1531, Sept. 1988.

[11] Laine U. K., Altosaar T., "An orthogonal set of frequency and amplitude modulated (FAM) functions for variable resolution signal analysis". Proc. of ICASSP-90, Vol. 3, pp. 1615-1618, Albuquerque, New Mexico, April 3-6,1990.

[12] Laine U. K., "FAM class of orthonormal functions, differential operators and integral transforms" (manus.), 1993.

[13] Laine U. K., "Famlet, to be or not to be a wavelet". IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis, Victoria, British Columbia, Canada, Oct. 4-6 1992.

[14] Laine U. K., "MSE filter design and spectrum parametrization by orthogonal FAM Transform". Proc. of the ISCAS-93, Chicago, Illinois, USA, 1, pp. 148-151, 1993.

[15] Saramäki T., Renfors M., "A novel approach for the design of IIR filters as a tapped cascaded interconnection of identical allpass subfilters", Proc. of ISCAS 1987, Philad., PA, May 1987.

[16] Steiglitz K., "A note on variable recursive digital filters", IEEE Tr. Ac, Speech and Signal Proc., ASSP-28(1), Feb. 1980.

[17]. Markel J. D., Gray A. H., *Linear Prediction of Speech*, Sringer, 1976.

[18] Altosaar T., Karjalainen M., "Diphone-based speech recognition using time-event neural networks". Proc. of ICSLP-92, Banff, Canada, 1992.

[19] Karjalainen M., "Auditory models for speech processing". Proc. of Int. Congr. of Phonetic Sciences, Tallinn, Estonia, 1987.