6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

VERY LOW BIT RATE VOICE CODER BASED ON A NONLINEAR HEARING MODEL

Rudolf Földvári, László Gyimesi

Széchenyi István College Győr,
Department of Automation,
H-9026 Győr, Hédervári u. 3., Hungary

10.21437/Eurospeech.1999-348

ABSTRACT

A new hearing model based on the knowledge of anatomy and phychoacoustical phenomena was presented by R. Földvári and Gy. Ács in 1996. After a short survey of older hearing models, they proved that their hearing model is able to describe true hearing experience (critical bandwidth, phase limit frequency). We suppose that the outputs of Zwicker's filters and the nonlinear transformation secured by the nonlinear hearing model is suitable for brain processing. Furthermore, we wish to point out that by realizing this method a very low bit rate vocoder and in addition an efficient speech enhancement can be achieved in cases when signal-to-noise ratio is about 0 dB.

INTRODUCTION

The most widely used time-frequency representation of signals are spectrograms obtained by the short-time Fourier transformation. Fourier transformation is the basic method for signal processing but non-stationary signals also contain time dependent spectral components. We are about to show that by the using a Zwicker's filter bank and the introduction of a nonlinear transformation, it is possible to model human hearing respectively to model "human signal processing".

The Structure of the Human Hearing Organ

The human ear has three distinct regions: the outer ear, the middle ear, and the inner ear (Fig. 1). The outer ear consists of the pinna, and the external canal. The role of the outer ear is small. Sound waves are transmitted through the outer ear to the middle ear, which consists of the ossicular chain (hammer, anvil, and stapes), and transmits the acoustical sound waves (mechanical vibrations) to the inner ear. The middle ear is separated from the outer air by the ear drum but the Eustachian tube connects it with the nasal cavity (and thus of course with the outer air; no "aer internus" does exist) [2]. The cochlea which is a chamber filled with fluid and partitioned by the basilar membrane can be found in the inner ear.

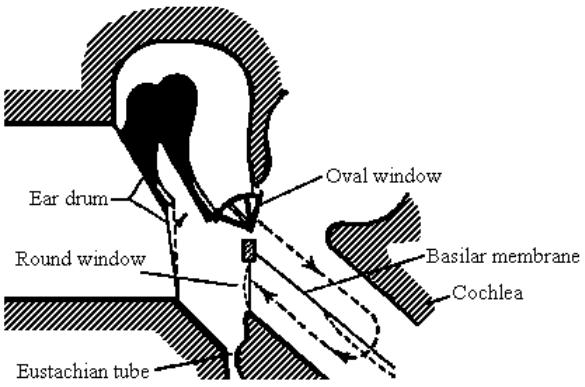

Fig. 1. The scheme of the hearing organ

On the basilar membrane, the organ of Corti is situated which can be seen in Fig. 2.

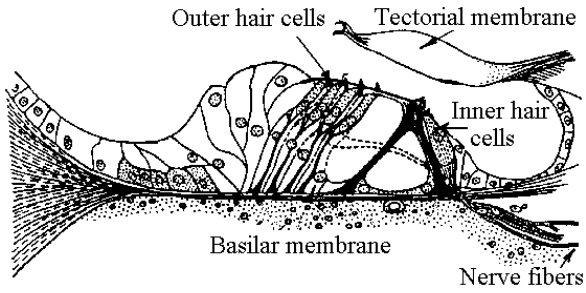

Fig. 2. The organ of Corti

Békésy (1899-1972) pointed out on experimental basis that the basilar membrane (together with the organ of Corti and the fluid of the cochlea) were too stiff to produce stationary waves [3]. Békésy proved that the mechanical vibrations of stapes on the oval window at entrance to the cochlea produce travelling waves. The basilar membrane shows a deflection maximum when a fixed frequency is applied but this maximum is flat. Békésy and other authors suppose that there must be a secondary filtering function in nerve fibres called lateral inhibition. Until today, however, no secondary filtering mechanism could be found. The basilar membrane vibrates together with the organ of Corti, in which there are four layers of inner hair cells and one layer of outer hair cells. The sensitiveness of the outer hair cells is by

40 dB lower than that of the inner hair cells. In the hair cells, mechano-electrochemical conversions are accomplished and in them start the afferent auditory nerve fibres to the brain.

Hearing Threshold and Masking Effects

If loudness is very small no sound at all can be heard. By increasing loudness, the sensation of sound suddenly appears in our minds. This sound level is called the hearing threshold (Dashed lines in Figs. 3 and 4). A loud sound makes the hearing of a weaker one impossible. This phenomenon is called the masking effect. The efficiency of masking depends on the frequency and the loudness of the masking tone (Fig. 3). If the masked and the masking sounds are conducted to the ears separately, the masking effect will seem much smaller because in such cases the masking effect originates in the inner ear and not in the brain. In the case of "white noise" we get a typical masking curve, which at approximately 1 kHz is by 17-18 dB higher than the spectral density of the masking noise; above 1 kHz the curve increases proportionally to frequency (Fig. 4).

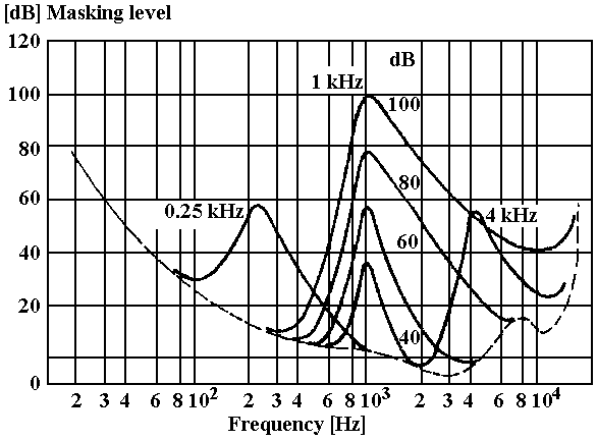

Fig. 3. Masking curves

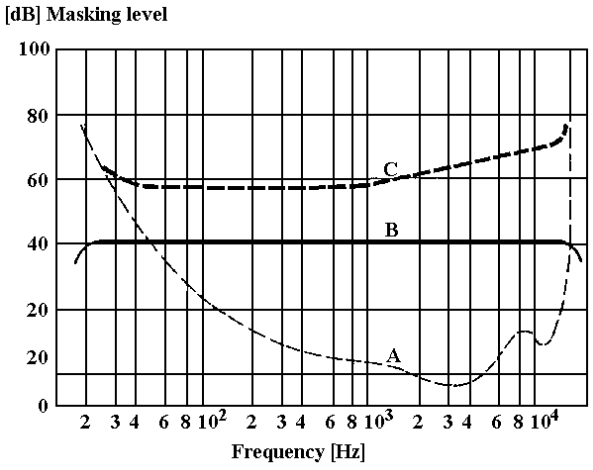

Fig 4. Masking curve in ”white noise”

Critical Bandwidth and Phase-Limit Frequency

Our sensitivity to loudness depends on the bandwidth of sound stimuli. Having crossed a limiting bandwidth value, the sensation of loudness will increase more than the intensity of the stimulating sound. Below a well-defined bandwidth, our hearing will reproduce loudness according to the sum of transmitted energy. If, however, the bandwidth overgrows a given value, the sensation of loudness will increase more than the physical intensity of the sound impulse. This limit is called the critical bandwidth. It was studied in detail by Zwicker [4].

Zwicker (1961) pointed out that critical bands have a certain width but their position on the frequency scale is not fixed. Thus a critical bandrate function was proposed by Zwicker [10] in graphical form. Zwicker and Terhadt suggested (1980) analytical expressions for critical bandrates and critical bandwidths. Later an improved 1-Bark bandwidth auditory filter was proposed by **Sekey** and **Hanson**. There are altogether 25 filters in the whole hearing region approximately 12-15 of them are in the region of speech. The slope of the filter in the stop-band was measured by Zwicker and Feldtkeller on the base of the masking effect [5].

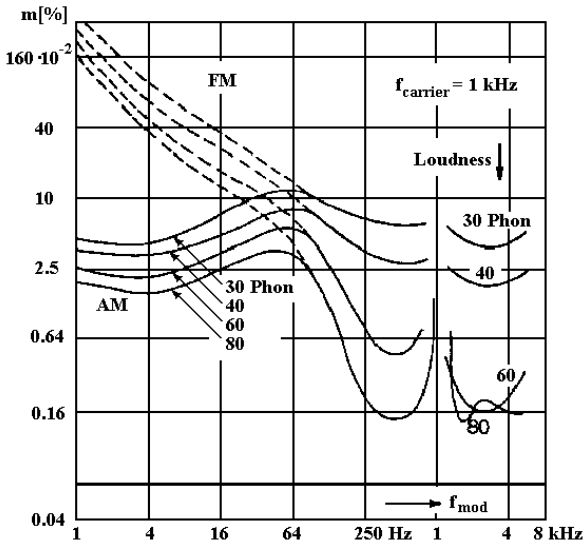

Fig. 5. The limit values of audible AM and FM modulations

Zwicker (1967) investigated not only critical bandwidths but also the limit values of audible AM and FM modulations [4]. He proved empirically that, within a certain bandwidth, our hearing is much more sensitive to AM than to FM modulation (Fig. 5). This phenomenon clearly shows that the same sound spectrum can be heard in different ways. The wide-spread statement that hearing is insensitive to phase is valid only without a well-defined bandwidth. Zwicker calls this bandwidth the phase-limit frequency. Although critical bandwidth and

phase-limit frequency have been defined differently, they proved to be equal in two entirely different psychoacoustical experiments. Thus they seem to be one universal constant.

Galambos and **Davis** (1943) investigated the response of single auditory nerve fibers to acoustic stimulation [1]. If a neuron cell was once active then, within an absolute refractory time, it is unable to give a second spike. The neuron cell has a relative refractory time when it is able to produce a new spike only at a higher level of stimulation. This phenomenon can well explain the idea of frequency/intensity connections. According to many researchers, the connection between a stimulating tone and the measured spikes below 4-5 kHz is synchronous. (The frequency of the highest fundamental tone of a pianoforte is 3520 Hz.) Using higher frequencies, the synchronous connection ends. If the stimulating tone has a higher level the travelling wave on the basilar membrane becomes wider, and more spikes are initiated in the brain.

The Suggested Nonlinear Hearing Model

On the ground of the facts described above, we suppose that all necessary facilities for working up auditive informations are at our brains' disposal. The filtering effect for the bandwidth of one third of an octave has developed on the basilar membrane, and the masking effect has developed, too. Although this model does not follow strictly the structure of the inner ear we think that it describes correctly the function of human hearing. The first part of our model consists of 25 Zwicker's filters which are given in [6],[7] and their characteristics can be seen in Fig. 6. After filtering, the signal is limited to a relatively narrow band. The second part of the model transforms the signal. The generalized amplitude function can be defined as

$$A(t) = x(t) \cdot \left(\cos\left(\Phi_0 + \int_{t_0}^{t} \Omega(\tau) d\tau \right) \right)^{-1}$$

where

$$\Omega(t) = \frac{d}{dt}\Phi(t) = \frac{x(t) \cdot y'(t) - x'(t) \cdot y(t)}{x^2(t) + y^2(t)}.$$

The x(t) is the output signal of the filter and its Hilbert pair is denoted by y(t). Henceforth this transformation shall be represented $\mathrm{Z}(\omega, t)$

$$\mathrm{Z}(\omega, t) = \begin{cases} A(t), & if \quad \omega = \Omega(t), \\ 0 & elsewhere, \end{cases}$$

and called the Generalized Amplitude and Frequency Transformation, **GAFT** [8]. It may thus be regarded as two mutually independent AM and FM demodulators. By applying this model, we get 25-25 slowly varying $A_k(t)$ and $\Omega_k(t)$ instantaneous parameters. This transform can be repeated arbitrarily many times because $A_k(t)$ and $\Omega_k(t)$ are equally analytic functions as $x_k(t)$ is. It can be proved that if a signal is a relatively narrow band then is not necessary to know its Hilbert pair for the determination of $A_k(t)$. Thus it is sufficient to calculate the filtered absolute value of $x_k(t)$ to find an approximate $A_k(t)$.

Altogether there are 25 filters in the whole hearing region; approximately 15 of them are in the region of speech. The periodicity of the high passed $A_k(t)$ gives the pitch (*or* fundamental) frequency and the low passed $A_k(t)$ gives the level in the channel. In cases of voiced parts, time functions have the same periodicity on each output of the Zwicker's filters. Of course, the filtered signals also have the same periodicity. A good way to determine this pitch frequency is to use a high pass filter and a frequency discriminator [9]. If, because of the influence of noise the signals differ, we estimate - out of these signals - a common pitch frequency. Thereafter the minimum and the maximum values of $A_k^L(t)$ in a 100 ms moving average are searched for to set the parameters of nonlinear processing [9].

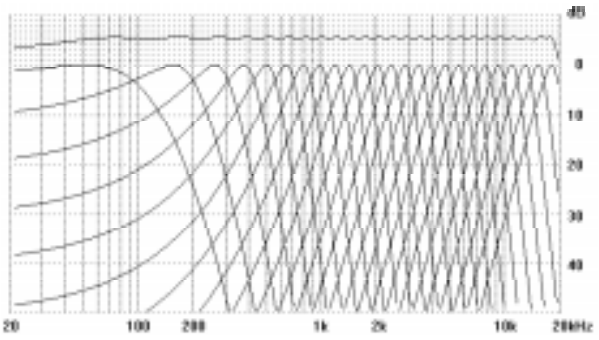

Fig 6. The characteristics of Zwicker's filters and the resultant of them

Practice has shown that noise is most disturbing in silence when speech is absent. We tried to cancel this noise in our speech enhancement technique. The model which can be seen in Fig. 7. illustrates our speech enhancement technique. It is evident that speech and noise signals together have greater intensity than has noise alone. The main philosophy of our method is: it is worth while to cancel the more silent parts and to enhance the louder parts. This, of course, can be made by a nonlinear processing. After processing the amplitude function, it will be easy to decide whether the signal was speech or noise. Thus we can decrease the influence of noise. This procedure is most efficient in speech silences. The parameters of nonlinear processing depend on the actual signal-to-noise ratio and these parameters must be updated during the whole time of processing. The sum of outputs of nonlinear processing is a noise-cancelled signal. According to experience, such speech is even more understandable than the original one. This procedure can decrease the influence of the coloured

noise signal too, because the channels set the features of nonlinear characteristics independently. Our method works successfully if the background noise is not a Gaussian process, even when it is only a disturbance, for example a sinusoidal or some other similar signal.

Simulations have shown that by using this hearing model we can achieve low data rate, too [10]. After calculating the frequency of $A_k^H(t)$ we get the pitch frequency which was linearly quantized with 8 bit and the $A_k^L(t)$ was logarithmically quantized with only 4 bit. The speech perfectly intelligible but the timbre of the speaker is not the same as the original one at the beginning.

The algorithm collects the features of timbre and after approximately one or two minutes "learning time" it is not possible to distinguish the input and output signals.

We wish to point out that by realizing this method we get, a data rate of 5 kbit/s, speech of full quality, same as a cassette tape recorder can provide. It is possible to achieve a data rate of 2 kbit/s the PCM quality. The simulations show that the highest achievable compression is about 1000 bits/sec which does not occur loss of PCM sound quality but in this case it is needed a longer, about five minutes, "learning time".

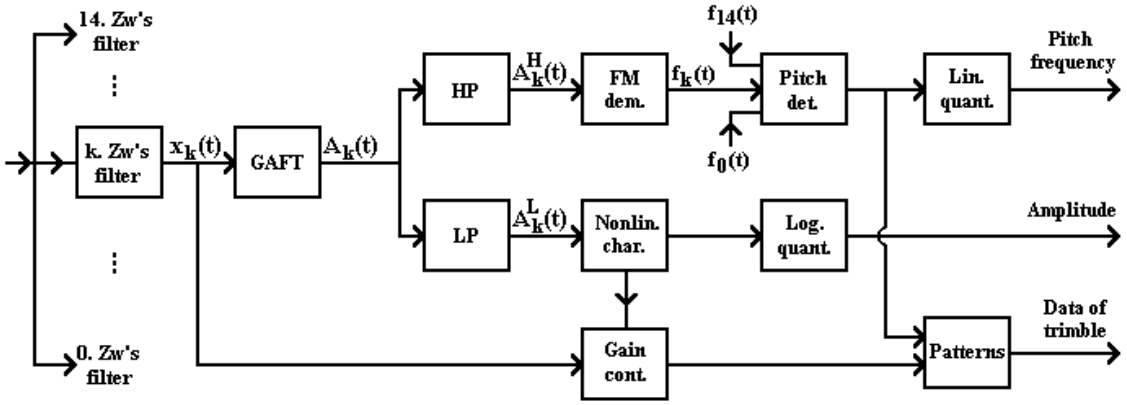

Fig 7. The block diagram of the suggested vocoder

CONCLUSION AND OUTLOOK

It has been found possible to model human hearing by the use of a Zwicker's filter bank and of instantaneous parameters. The model describes correctly the phenomena of phychoacoustical experiments, the phenomena of the musical ear included. By applying a simplified model we get a very efficient enhancement method and a low bit rate vocoder which is able to work even at a 0 dB signal-to-noise ratio.

REFERENCES

[1] Galambos, R., Davis, H.: *The response of single auditory nerve fibers to acoustic stimulation.* J. Neurophys., Vol. 6, 39-57, 1943.

[2] Békésy, Gy., Rosenblith, W. A.: *The early history of hearing. Observations and theories*. J. Acoust. Soc. Am. Vol 20, 1948.

[3] G. v. Békésy: *Experiments in hearing.* McGraw-Hill, New York, 1960.

[4] E.Zwicker, R.Feldtkeller: *Das Ohr als Nachrichtenempfänger.* Hirzel V., Stuttgart, 1967.

[5] E.Zwicker, E.Terhardt (editors): *Fact and Models in Hearing.* Springer V., Berlin - Heidelberg - New York, 1974.

[6] E.Zwicker, E.Terhardt: *Analytical Expressions for Critical Bandrate and Critical Bandwidth as a Function of Frequency* Acoust. Soc. Am. Vol. 68, 1523-1525, 1980.

[7] A.Sekey, B.A.Hanson: *Improved 1-Bark Bandwidth Auditory Filter.* Acoust. Soc. Am., Vol. 75, No. 6, 1984.

[8] R. Földvári: *Generalized instantaneous amplitude and frequency functions and their application for pitch frequency determination.* Journal of Circuits, Systems, and Computers, Vol. 5, No. 2, 1995.

[9] R. Földvári, Gy. Ács: *Speech Enhancement Based on a New Hearing Model.* 19th Czech-Hungarian-Polish Workshop on Circuit Theory and Applications, Prague, 1966.

[10] R. Földvári, Gy. Ács: *Speech and Music Coder Based on a New Hearing Model.* 7th Conference and Exhibition on Television and Sound Technique, Budapest, 1996.