A BINAURAL AUDITORY MODEL FOR SOUND QUALITY MEASUREMENTS AND SPATIAL HEARING STUDIES

Matti Karjalainen

Helsinki University of Technology
Laboratory of Acoustics and Audio Signal Processing
Otakaari 5A, FIN-02150 Espoo, Finland

ABSTRACT

This paper addresses auditory modeling in relation to binaural applications. Relatively few studies so far exist where auditory modeling is extended to include advanced binaural features. There are numerous potential applications where such modeling may yield practical results or improved insight into the problem. The focus here has been to develop models for spatial sound quality measurements and general modeling of spatial hearing. The novelties of the paper are related to auditory filterbank designs, temporal processing schemes, and binaural processing strategies.

1. INTRODUCTION

All acoustic and audio systems require perceptual evaluation of results or guidance during the design process. In most cases this is done by informal listening, possibly supported by objective measurements. In some cases a carefully controlled subjective listening experiment is necessary, which is the best way to obtain a relevant picture of sound characteristics, but it is a laborious task and the repeatability of results may not be good. As an increasingly attractive alternative, computational auditory models are used to combine the advantages of subjective listening with the robustness and ease of objective measurements.

Computational modeling of the auditory system has been applied successfully to demanding problems in speech and audio system design and sound quality measurements, e.g., [1], [2], [3], [4]. In fact, audio compression schemes such as the MPEG standard [5] are based on implicit or explicit psychoacoustic models and are a part of the encoder-decoder process. Computationally or conceptually, auditory models have also been used in noise evaluation [6], concert hall acoustics, modeling of binaural hearing [7], [8], and in many speech recognition systems such as [18], etc.

Continuing efforts are being made to develop auditory models in order to understand human hearing and to find potential applications. Of the two main approaches to auditory modeling, the physiological and the psychoacoustical approach, the latter one is normally more attractive in many engineering applications since it abstracts from the innumerable known and unknown details of human hearing and concentrates on observable behavior, i.e., input-output relationships.

In this paper we present a psychoacoustically oriented auditory model which has been developed for a variety of applications that require relatively detailed time-frequency characteristics, binaural capability, and computational efficiency for real-time or near real-time processing on modern signal and RISC processors. We first give a general description of the model and then present the subunits, their properties and implementation. The main novelties are: (a) warped filter design of critical-band filterbanks, (b) feedback-based realization of a temporal dynamics processor for complementary fast (firing rate) and slow (loudness spectrum) response outputs, and (c) a binaural feature analysis scheme. The properties of the model will be compared with the properties of the human auditory system. Finally, we will discuss how to adapt the model to various applications. We are assessing the feasibility of a real-time version of the model on a multi-DSP system.

2. STRUCTURE OF THE MODEL

The binaural auditory model consists of three main parts (see Fig. 1): (a) auditory filterbanks for critical-band bandpass filtering, (b) temporal dynamics processors for firing rate and loudness analysis, and (c) a binaural feature analyzer that combines information from both channels.

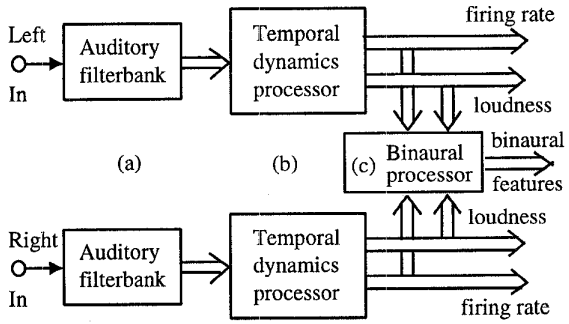

Fig. 1 *Block diagram of the main processing units in the binaural auditory model.*

3. AUDITORY FILTERBANKS

There exist many methods to approximate the frequency resolution of the auditory system by different selectivity curves or filter responses; for example the Bark scale and the spreading pattern defined by Zwicker [9] and expressed as a function by Schroeder et al. [1], the cascaded filter structures by Lyon [10], and the gammatone filterbank by Patterson and Holdsworth for the ERB-rate scale [11]. The problem in our case has been to design low order, robust filters that match the psychoacoustic frequency response of hypothetical channels in the human auditory system.

We have used auditory filterbanks that split the incoming signal into frequency bands of 1 Bark each, i.e., critical bands

0-7803-3192-3/96 $5.00©1996 IEEE

[9]. A varying number of channels may be used; 24 Bark channels is the minimum number to cover the audio range but more (overlapping) channels may be used if an application so requires. We have applied the following three filter designs.

The first two techniques are based on warped filter design: warped FIR (Fig. 2a) and warped IIR (Fig 2.b). Variations of the warping idea have been studied, e.g., by Oppenheim et al. [12], Strube [13], and Smith [14]. Laine has developed further generalizations based on FAM functions [15].

Fig. 2a is called a warped FIR (WFIR) since it is like an FIR filter but uses allpass sections $(z^{-1}-\lambda)/(1-\lambda z^{-1})$ instead of unit delays. (In fact, the structure is recursive and thus IIR from traditional filter design point of view). With a proper value of λ the warped frequency scale fits well to the Bark scale: for example $\lambda = 0.63$ is used for a sampling rate of 22 kHz. Smith and Abel [16] have derived a formula for an optimal match as a function of the sampling frequency.

The structure in Fig. 2a has attractive properties compared to normal FIR filters: it can be designed directly in the Bark scale domain, it has minimal precision requirements, and it reduces the order of the lowest Bark channel filters radically, e.g., from about 1000 to 64 taps for a sampling rate of 22 kHz and selectivity curves of [1] with a 100 dB dynamic scale. Since all filters of the bank share the same allpass delays, the increase in computational cost due to allpass delays remains relatively small.

Fig. 2b shows a new realizable warped IIR filter (WIIR) that is the second alternative for an auditory filterbank that we have considered. It shares some of the advantages of warped filters but due to a relatively complex structure and IIR design it was not found as attractive as the warped FIR for this application.

A more detailed analysis of warped filter designs will be given elsewhere [17].

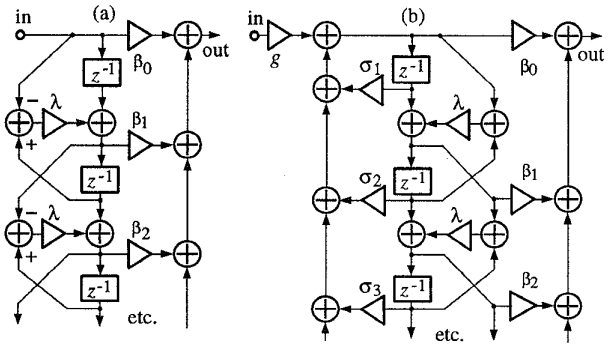

Fig. 2 *Warped filter structures for auditory filterbanks: (a) warped FIR (b) warped IIR structure.*

The third filterbank technique that we have used is similar to the parallel-cascade form used by Lyon [10] except that the properties of the structure are tuned to psychoacoustic data rather than physiological features of the basilar membrane. This is the computationally fastest implementation for an auditory filterbank, e.g., for a real-time version, although there is not as much freedom to tune the filter responses as with the structure shown in Fig. 2a. A total filter order of about 100–150 for a single ear filterbank, including downsampling for the lowest Bark channels, represents a very efficient implementation.

4. TEMPORAL DYNAMICS PROCESSOR

An auditory model must simulate the temporal behavior of human hearing, including pre- and postmasking, temporal integration [9], as well as various adaptation phenomena on the auditory nerve level [18]. This part of a model must also be inherently nonlinear: the hair cells detect the excitation level by half-wave rectification and the time constants can not be simulated by linear filters only. The temporal processing must be detailed enough in order to allow for accurate analysis of binaural features, such as interaural time differences. Block-based analysis using FFT and spectral warping is not suitable for this purpose.

In our model (see Fig. 3) the filterbank channels are followed by half-wave rectification and some low-pass filtering to simulate the loss of synchrony of nerve firings at high frequencies. Downsampling may be applied after the lowpass stage to reduce the computational load per channel for the next stages if savings for a real-time version are needed. This is not desirable, however, if high binaural accuracy is needed.

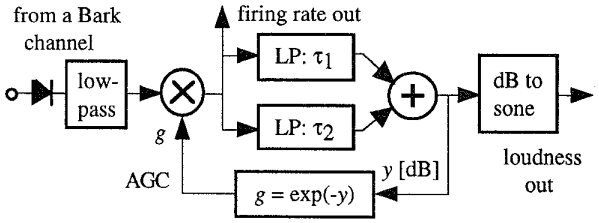

Fig. 3 *Temporal dynamics processor of a single channel.*

The temporal dynamics processing is realized by an automatic gain feedback (AGC) system as shown in Fig. 3 for a single channel. The complementary nature of the firing pattern at the auditory nerve as a highpass response and the loudness output as a lowpass response draws attention to this possibility. As shown in the step responses of Fig. 4, proper exponential gain feedback and temporal integration filters in the forward part of the loop in Fig. 3 simulates this well.

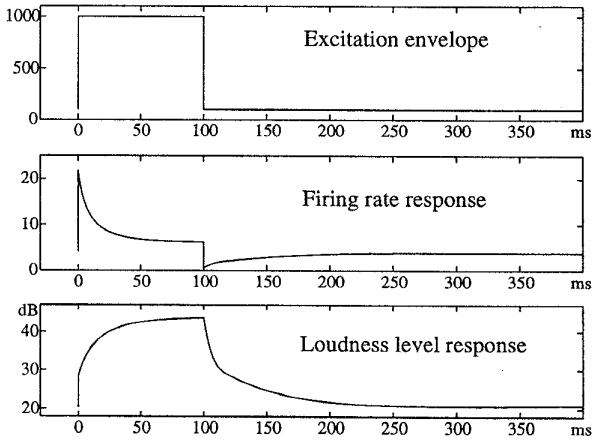

Fig. 4 *Firing rate and loudness responses of the temporal dynamics processor to an excitation having a step function envelope.*

The firing pattern shows a strong overshoot on attacks and adaptation-like decay thereafter. The loudness level output grows in a way that matches the loudness formation rule of short sounds [9]. When the excitation is stopped, the loudness level output matches relatively well to known postmasking (forward masking) patterns [9]. In order to achieve these behaviors, a second-order filter (two parallel first-order filters in Fig. 3) is needed in the forward path with time constants $\tau 1 \approx 130$ ms and $\tau 2 \approx 10$ ms for falling excitations and $\tau 1 \approx 1.0$ s and $\tau 2 \approx 100$ ms for rising excitations.

A technical trick, a feedforward path (not shown in Fig. 3), is further needed to limit the AGC gain for fast onsets. This is partly due to the discrete-time implementation where a unit delay is the minimum feedback period instead of practically immediate feedback in a continuous-time system.

Finally, in the temporal processor of Fig. 3, the logarithmic output y is mapped to the specific loudness (sones) of the channel.

5. BINAURAL PROCESSOR

The third subsystem of our auditory model (Fig. 1), the binaural processor, combines information from the temporal dynamics processors of the left and right ear models into features of interest in binaural applications.

An extensive body of knowledge related to spatial hearing and sound localization is compiled in [19] and [20]. Due to the complexity of the problem domain, there exist relatively few computational auditory models with advanced binaural features. It is well known that the two main primary cues to be included are the interaural time (ITD) and level differences (ILD), but these are not enough since, e.g., sound sources in the vertical median plane do not produce any ITD or ILD [20, p. 15]. If the interaural cross-correlation (IACC) [19, p. 255] is used instead of simple ITD, there is enough information for accurate estimation of the elevation, as is clearly demonstrated in [8]. The IACC and ILD analysis may also be combined [21], [22].

The *precedence effect* [23] must be included in a binaural model especially when room reflections and reverberation are present in the signals to be analyzed. Examples of such computational models are given, e.g., in [21] and [24].

We proposed earlier a binaural model for stationary signals which was shown to be very precise in estimating the azimuth and elevation of a sound source in simple acoustical conditions [8], outperforming human abilities. In the present new model we follow the same approach but add dynamic features, including the precedence effect. It is natural to apply binaural processing to the firing rate responses of the temporal processor (Fig. 3). A combined loudness can be computed from the left and right channel loudnesses, although this information is not used in the present binaural processor.

Figure 5 shows the basic configuration of the binaural processor, which computes two basic features for each Bark channel pair as used in [8] but now in a more dynamic way: (a) running IACC of the firing rate signals and (b) level-ratio of the same signals. The level ratio R is defined as

$$R = (Fr - Fl) / (Fr + Fl),$$

instead of a level difference in dB, since its value is always bound between -1 and +1. Fr and Fl are the firing rate outputs of the right and left channels, respectively.

The IACC is computed as a running short-time cross-correlation between the firing rate outputs of the right and left channels with a 5...10 ms decay time constant and a $\pm$ 1 ms time window. This is done for all critical band pairs, thus it is a very time-consuming process where optimizations are highly needed. Since the firing rate outputs of the temporal processor are lowpass filtered, they may be decimated by a factor of 2 to 3 without too much error, in order to reduce the computational load of cross-correlations.

Notice that the IACC and level ratio processing are applied to the firing rate signals that have a strong overshoot of onsets. Thus they automatically approximate the precedence effect to some degree, i.e., they are sensitive to primary onsets and less sensitive to delayed versions of the signal arriving within a short period (such as early reflections from the walls in a room). Further realization of the precedence effect relies on the following stages of the binaural processor.

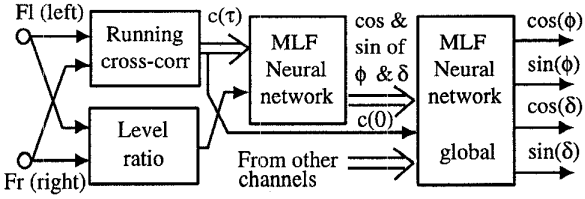

Fig. 5. *Block diagram of the binaural processor.*

The primary binaural features, i.e., the IACC and ILD, of each Bark channel are mapped to secondary features, the sines and cosines of the angle estimates for azimuth (ϕ) and elevation (δ). This is done by multilayer feedforward neural nets (MLF) that are trained from examples in a supervised manner, in a way similar to the earlier model [8], (cf. [24], [25]), but now dynamically in time. Such a neural network approach is a natural choice since it can aquire complex nonlinear mappings and properties of the specific front-end of the model. For a review of recent connectionist models, see [26].

The directional estimates from the neural net of a Bark channel pair are further combined using a global directional analyzer, which is also an MLF neural network. It is trained as a second phase after the channel-specific nets have been trained. A typical dimensionality of the channel-specific net is 2x(23)+1 inputs, 4 hidden nodes, and 4 output nodes. The global network has 24x(4+1) inputs, 4 hidden nodes, and 4 output nodes. The computation rate of neural networks is once every 2...5 milliseconds, based on down-sampled versions of the IACCs and ILDs.

The neural networks shown in Fig. 5 do not include any temporal memory. Thus they do only instantaneous estimation of the directional parameters. Probably a better method might be to use recurrent networks that can learn temporal behavior [27, p. 20]. An interesting alternative for the processor of Fig. 5 is to replace the entire system with a single recurrent network with inputs from all firing rate signals of the temporal processors. The disadvantage is that then no channel-specific directional parameters are available.

Since the cross-correlation analyses for Bark channel pairs produce much more data than is in the original signals, data reduction is preferable in some applications. Some possibilities are: (a) more smoothing and downsampling in the time dimension, (b) merging of Bark channels by averaging [8], or (c) reducing each cross-correlation to a single number such as a peak position or centroid as estimates of the interaural time difference. The last one is undesirable if the elevation estimate is needed, since in the median plane no cues are retained for the task.

6. APPLICATION STRATEGIES

The binaural auditory model described above is designed for a set of potential applications, such as modeling of directional hearing [8] (source location: direction, distance), measurement and evaluation of acoustic spaces (rooms, concert halls), sound reproduction (HiFi and PA loudspeakers, 3D and surround sound systems), stereo audio codecs, noise quality, speech recognition in a noisy room, etc.

In a typical case the model is used together with a dummy head or insert microphones. These left and right ear signals are fed as inputs to the filterbanks. In the case of computational modeling of a room or hall, the auditory model inputs may be obtained using measured head-related transfer functions or corresponding impulse responses instead of a dummy head.

Different applications require different outputs from the model. In some objective measurements even the raw outputs of the filtebank or the temporal processor are interesting. Often the primary binaural cues (ITDs and ILDs) are a natural choice. The precedence effect could also be modeled separately for measuring the noticeability or prominence of reflections and echos. The binaural processor of Fig. 5 is suitable for applications where the direction (azimuth and elevation) of a source must be estimated.

In an important group of applications the output is a parameter or a set of parameters that has a complex and nonlinear relation to the acoustic properties of the incoming sound. If the mapping from primary binaural ques is not too complex, a multidimensional lookup table and interpolation may be used to obtain the final features or parameters. A general-purpose method to be tried in complex cases is the use of neural networks, as proposed above. Even relatively vague subjective parameters, such as sound quality measures or preferences, may be useful as far as two prerequisites are met: there is enough data on the desired input-output relationships and a neural network configuration is found that is able to learn this relationship accurately enough.

7. SUMMARY

Our deterministically formulated binaural auditory model has been tuned to match the major psychoacoustic properties of the human auditory system relatively well. Among these properties are: critical bands and masking in the frequency domain, time domain behavior including pre- and postmasking as well as temporal integration, good temporal resolution at the firing rate level, and at least qualitatively correct binaural feature analysis. Models with such features are highly desirable as tools for basic research and development of binaural technology.

First experiments show that, due to inclusion of dynamic features, the model performs better sound source direction estimation in reverberant and noisy environments than our previous "steady-state" model [8]. Further work is needed to fine-tune the model and its subunits according to psychoacoustic data known from the human auditory system, and to show the match between these behaviors in cases that are important theoretically or from application points of view.

There exist also several variations to the binaural processing scheme to be experimented with. One of the important efforts is to optimize the DSP algorithms and implement the model on a fast processor or a multiprocessor, in order to make it more feasible in practical applications. Now the present model takes about 12 times real time on a 33 MHz TMS320C30 with a sampling rate of 22 kHz (signal bandwidth is 10 kHz). Many applications, such as measurements with predefined test signals, allow for compromising some features of the model. This may radically simplify or speed up the computations.

ACKNOWLEDGMENT

This work has been supported by the Academy of Finland.

REFERENCES

[1] Schroeder & al., "Optimizing Speech Coders...," JASA 66(1979).
[2] Karjalainen, "A New Auditory Model...," ICASSP-85.
[3] Brandenburg, "Evaluation of Quality for ...," 82nd AES Convention, preprint 2433, London 1987.
[4] Beerends and Stemerdink, "A Perceptual Audio ...", JAES vol. 40, No 12, 1992.
[5] Brandenburg, "ISO/MPEG-Audio Codec," 92nd AES Convention, Vienna, 1992.
[6] Fastl, "Psychoacoustics and Noise Evaluation," in NAM94 Proceedings, Aarhus, Denmark 1994.
[7] Lyon, "A Computational Model of ...," ICASSP-95.
[8] Backman and Karjalainen, "Modelling of Human ...," ICASSP-93.
[9] Zwicker and Fastl, *Psychoacoustics.* Springer 1990.
[10] Slaney, "Lyon's Cochlear Model," Apple Technical Report #13, 1988.
[11] Patterson & al., "Complex Sounds and Auditory Images," in *Auditory Physiology and Perception* (ed. Cazals & al.), Pergamon 1992.
[12] Oppenheim and Johnson, "Discrete Representation of Signals," *Proc. IEEE*, vol. 60, 1972.
[13] Strube, "Linear Prediction on a Warped Frequency Scale," JASA vol. 68, nr. 4, Oct 1980.
[14] Smith, *Techniques for Digital Filter ...* Report STAN-M-14, CCRMA Stanford, 1983.
[15] Laine et al., "Warped Linear Prediction (WLP) in Speech and Audio Processing," ICASSP-94.
[16] Smith and Abel, "The Bark Bilinear Transform", IEEE ASSP Workshop, Mohonk, New Paltz, 1995.
[17] Karjalainen & al., Unpublished manuscript, HUT Acoustics laboratory.
[18] Seneff, "A Joint Synchrony/Mean-Rate Model of ...", Journal of Phonetics, vol. 16, 1988.
[19] Blauert, *Spatial Hearing.* MIT Press, 1983.
[20] Yost and Gourevitch (ed.), *Directional Hearing.* Springer Verlag, 1987.
[21] Lindemann, "Extensions of a Binaural..., I & II.", JASA 80(6), 1986.
[22] Bodden, "Modeling Human Sound-Source Localization ...", Acta Acustica, 1:43-55,1993
[23] Zurek, "The Precedence Effect", in [20].
[24] Martin, *A Computational Model of ....* M.Sc. Thesis, MIT, 1995.
[25] Lim and Duda, "Estimating the Azimuth and Elevation of a Sound Source ...", 28th Asilomar Conf. on Signals, systems, and Computers, 1994.
[26] Duda, "Connectionist Models for Auditory Scene Analysis", in *Advances of Neural Information Processing*, Morgan Kaufman Publ. Inc., 1994.
[27] Haykin, *Neural Networks*, IEEE Press, 1994.