IEEE 1995 Workshop on Applications of Signal Processing to Audio and Acoustics, Mohonk Mountain House, New Paltz, NY

Binaural Modeling and Auditory Scene Analysis

Markus Bodden

Lehrstuhl für allg. Elektrotechnik und Akustik (Prof. Dr.-Ing. Dr. techn. hc. Jens Blauert),
Ruhr-Universität Bochum, D-44780 Bochum, Germany

ABSTRACT

During the past years auditory scene analysis has become a focus of research on the human auditory system (Bregman, [1]). The abilities of humans to analyze their auditory environment are so striking that they are attracting researchers from different areas. Traditionally, research on this topic was initiated and mainly performed by psychologists. But, besides the main interest to understand the underlying mechanisms, engineers are getting more and more involved. The reason for this is that possibilities to simulate auditory scene analysis offer new solutions to a variety of technical problems. Due to that, the model presented in this article, which gives a review of the work performed at the institute of acoustics of Bochum University in Germany, is mainly motivated by this problem-oriented thinking.

In contrast to the majority of activities in the field of auditory scene analysis our approach is based on binaural peripheral processing. The intention to start this work more than ten years ago was to understand and simulate binaural hearing. As a result of this work several models of the binaural auditory system have been developed, ranging from closely physiologically oriented models to rather signal-processing-motivated algorithms.

Recently the focus of work has been shifted to applications of these models to technical problems. These applications can be grouped into two topics: *sound source localization* and simulation of the *Cocktail-Party-Effect*. This article is intended to describe a current model, show its applications, and to discuss its relation to auditory scene analysis.

1. INTRODUCTION

The performance of the auditory system with regard to the ability to analyze the auditory environment, the localization of sound sources, and the perception of speech is indeed striking. Even though this performance is yet hard to understand and simulate for single sound sources, the abilities of the auditory system go far beyond that: auditory streams can be grouped, sources localized, and speech be understood even in acoustically adverse conditions, that is, with several concurrent sound sources and interfering reverberation. From the point of view of theoretical signal processing it is striking that the auditory system can achieve all this with just two receivers, the left and the right ear.

As a consequence, these properties of the human auditory system have been a motivation to simulate it by means of computer models. On one hand these models serve as a tool to understand the underlying processing, on the other hand they can be applied to a variety of technical problems:

- *Localization models* can be used to find and identify sound sources, e.g., for advanced *noise measurement tools* (Bodden, [2]; Genuit and Blauert, [3]);
- *Sound source separation models* can be used for
  - (a) *speech recognizers* in order to maintain high recognition rates in the typical noisy applications scenarios, such as offices and public places, where traditional methods show a significant decrease of performance;
  - (b) *hearing impaired* persons who must come to terms with reduced intelligibility in noisy environments. Conventional hearing-aids do not offer much help with respect to this task;
  - (c) *hands-free telephony*, where a desired speaker is to be picked out of number of source signals whilst avoiding audible feed-back in the transmission chain;
- *Auditory Models* which are sophisticated enough to determine properties of hearing events can be used as a basis for methods to evaluate *Sound Quality* and tools for *Sound Design*. The latter two points have become very popular during the last years and are of high industrial relevance for all types of products emitting sound.

Up to today, we mainly concentrated on the development of binaural processing strategies. As a result, several models with different degrees of complexity and different basic approaches have been developed. Nevertheless, some rudimentary principles of auditory scene analysis are yet included, and for the actually planned extensions of our models knowledge about auditory scene analysis plays a crucial role.

2. THE BINAURAL AUDITORY MODEL

In this article we will concentrate on a binaural model developed at Bochum University. General overviews on models of binaural perception can be found in Colburn & Durlach [4], Blauert [5], and Stern [6].

A global structure of a binaural auditory model as proposed by Blauert [7] is depicted in Fig. 1. In the following section the current model will be described.

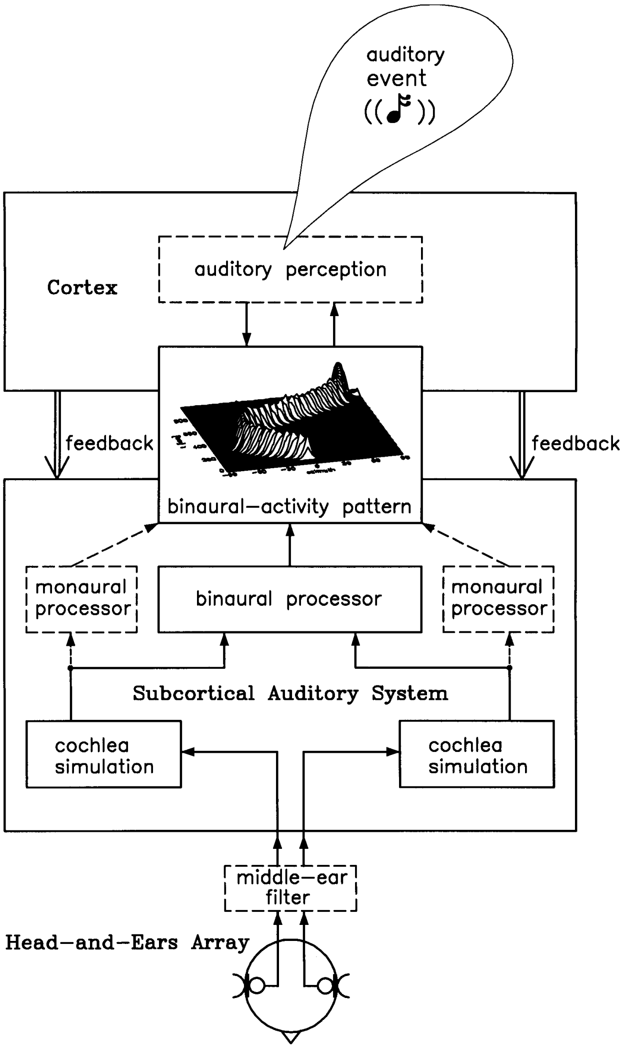

Figure 1: Global structure of an auditory model (taken from Blauert, [7])

2.1. Peripheral Stages

The peripheral block of the model comprises simulations of the outer, middle, and inner ears. The *outer ears* can be described in form of directional dependent transfer functions, the HRTF's (Head-Related Transfer Functions). They introduce interaural differences in time (IDT) and intensity (IID) into the signals that can be recorded at the eardrums of a listener. The *middle ears* are usually neglected since their transfer functions do not depend on the direction of sound incidence. The models for the *inner ears* can be rather simple for signal processing purposes. We use a bandpass filterbank with 24 filters corresponding to the bandwidth of critical bands as proposed by Zwicker and Feldtkeller [8], a halfwave rectification, a square root function, and a lowpass filter with a cutoff frequency of 800 Hz to extract the envelope at high frequencies. This process calculates probabilities of the firing of hair cells and thus leads to a deterministic model.

Lindemann [9] developed the binaural processor which is based on the idea of performing a running interaural crosscorrelation in the time domain. The bandfiltered signals from the left and right ear move in opposite directions along delay lines. The contents of the delay lines are multiplied at each tap and a running integration is performed. In contrast to other crosscorrelation models, Lindemann implemented a powerful inhibition mechanism directly into the correlation itself: *contralateral inhibition*. This inhibition is implemented by additional multipliers for each tap on the delay lines, resulting in time varying attenuations of the signals moving along the delay lines. The amount of attenuation is controlled by the amplitude of the signal at the corresponding tap of the contralateral delay line. The contralateral inhibition performs a *disambiguation* of the resulting crosscorrelation, a *contrast enhancement* of the correlation patterns and makes the correlation *sensitive to interaural intensity differences*. The latter point is due to the fact that contralateral inhibition becomes unsymmetrical if an interaural intensity difference occurs.

This processing scheme thus evaluates IDT's and IID's in a combined manner. Therefore Gaik [10] extended the model by an adaptation to individual HRTF's. He analyzed catalogues of HRTF's measured by Pösselt et al. [11] and calculated the interaural differences for 122 directions of the upper hemisphere in each critical band. As a result he was able to determine *natural combinations* of interaural differences. The adaptation process is implemented by an additional weighting of the signals moving along the delay lines. The weighting is performed in such a manner that contralateral inhibition becomes symmetrical for natural combinations of interaural parameters. The weighting coefficients are determined in a supervised learning phase from the natural combinations of interaural parameters.

Monaural processors have been added to the end of the delay lines to consider pure monaural hearing events, which can for example occur in headphone presentation.

In addition to the model described above different models have been developed for special purposes:

- a *probabilistic model* has been implemented by Wolf [12]. This model applies the same processing strategy to the output of probabilistic hair cell models. It thus shows a closer relation to physiology, but the performance is not superior to that of the model described above, although it is much more computationally expensive;
- a *simplified model* for signal processing purposes was presented by Grabke and Blauert [13]. By means of a peak detection mechanism the correlation is only performed for signal peaks, so that the computational effort has been significantly reduced by maintaining the main relevant features;
- a *statistical model* has been proposed by Slatky [14]. This model is based on a completely different approach. It evaluates the statistical properties of the interaural differences;

- a *frequency-domain model* currently is under development. This model is a pure signal processing tool and employs highly abstracted processing mechanisms.

2.2. Higher Stages

Simulations of higher stages of the auditory system and for sound source separation have been added to the model by Bodden [15]. The spatial distribution of sounds surely is one of the basic features contributing to the analysis of the auditory environment. Therefore the higher stages of the model show the close relation to auditory scene analysis. They aim at predicting sound source localization, separation of concurrent sources and evaluation of properties of hearing events.

The model is able to reproduce major effects of sound source localization. Besides fused hearing events resulting from head-related signals unnatural situations like headphone representations can be reproduced. Gaik and Wolf [16] showed for example that the model is able to predict the breakup of the hearing event into multiple images if the interaural differences are contradictory, as it can occur in trading experiments. Besides other effects, monaural events, summing localization, binaural beats, and the reduced localization accuracy with increasing azimuth can be reproduced.

The prediction of localization described here is restricted to the localization in the frontal horizontal plane. Front-back discrimination and elevation perception still are unsolved problems tackled by actual investigations (e.g., Hartung, [17]). Although signal processing tools like neural networks trained on the output of the binaural model can achieve correct localization rates of more than 80 %, the underlying mechanisms are not yet well understood.

If only one sound source is present in non-reverberant environment, the direction of sound incidence can directly be determined from the neural excitation patterns produced by the binaural processor. If more than one sound source or reflections are present, the auditory scene has to be analyzed in order to get stable results. Therefore the following steps are performed:

- a *correlation-azimuth transformation* replaces the correlation axis representing the interaural delay with an axis representing the azimuth. The transformation rule is determined in a supervised learning phase from the output of the model itself for known directions of sound incidence;

- a *running average* using is applied using a time constant from 1 to 100 ms to smooth the neural excitation patterns;

- a *combination* of the information provided by each critical band is performed. A total neural excitation is determined as the weighted sum of the neural excitation of each critical band, simulating the combination across frequency in a simple way. The weighting function is again determined in a supervised learning phase from the output of the model itself by means of a comparison of the original azimuths to the predicted azimuths;

- a *determination of azimuths* builds the final stage. The onsets of neural excitations are analyzed and thresholding is considered to make up a decision whether a new hearing event occurred.

Fig. 2 shows an example of the output of the binaural processor for a single sound source and the predicted azimuths for two alternating moving sound sources.

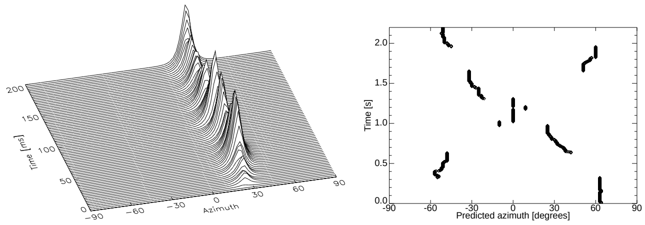

Figure 2: Left: output of the binaural processor, one sound source (bandfiltered noise, 400-510 Hz). Right: predicted azimuth for two alternating, moving sound sources.

2.3. Sound Separation

The neural excitation patterns produced by binaural models offer a decisive advantage: besides the dimensions time, frequency and amplitude an additional dimension is available - the spatial distribution of neural excitation. Using this information competitive signals from different directions of sound incidence can be separated: their respective excitation will be located at different positions in the excitation patterns (Bodden, [15]). Thus the so-called *desired excitation* (the neural excitation belonging to the desired source) is determined by windowing the binaural excitation with a window centered around the position of the desired speaker. The ratio of this desired excitation to the total excitation gives a measure of the actual signal to noise ratio, and represents a weighting factor which can be applied to the time signal of the respective critical band. Since these weighting factors are calculated as a function of time independently in each critical band, a time-varying transfer function is calculated. The weightings are applied to the critical band signals provided from the peripheral bandpass filterbank, and the sum forms the output signal. Either a one-channel signal (processing the signal with the better signal-to-noise ratio) or a binaural signal (processing both channels) can be produced.

Due to the high degree of nonlinear processing, i.e., contralateral inhibition, sharp beamforming can be achieved even at very low frequencies. The performance of the system has been evaluated by comprehensibility tests with hearing impaired subjects. Words (consonant-vowel-consonant clusters) and sentences of competitive speakers (2 male or 3 male/female speakers) have been mixed at a signal-to-noise ratio of 0 dB and presented to the subjects either in the original distorted version (dichotic) or in the processed version (diotic) via headphones. Fig. 3 shows the results. A significant gain in comprehensibility ranging from 12 to 34 % can be observed. It can be remarked that comprehensibilities for the consonant clusters are enhanced in the same range as for the vowel clusters.

2.4. Automatic Speech Recognition

The model described above was combined with a simple speech recognizer, a self-organizing feature map, in order to show the principle advantage of binaural processing for recognition in noise (Bodden and Anderson, [18]). Instead of using the model to pro-

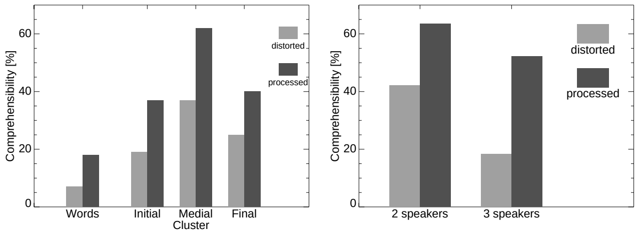

Figure 3: Result of the comprehensibility tests. Left: word test. Right: sentence test. Average of five hearing impaired subjects.

duce an enhanced signal to be fed into the recognizer an interface module was developed to extract a vector containing the desired neural excitation from the binaural excitation patterns. This module picks out the excitation corresponding to the direction of incidence of the desired speaker. The results of recognition experiments for a desired speech signal from the front and noise at 30^o are summarized in table 1 (speaker-independent and context-independent phoneme recognition rates). The results show that the binaural representation provides over a 20 dB SNR advantage over the monaural representation.

	Measure	SNR/dB						
		clean	21	9	3	-3	-9	-21
mon.	Correct	43.4	40.1	34.7	31.1	28.2	23.0	22.6
	Deletions	7.7	13.5	22.6	30.3	38.3	70.0	76.3
	broad class	72.3	66.5	60.2	56.2	50.6	26.6	22.9
bin.	Correct	43.4	42.7	41.0	39.3	37.7	36.1	30.4
	Deletions	7.8	8.2	10.0	11.8	13.9	18.1	32.3
	broad class	71.9	71.6	69.7	67.9	65.7	62.7	55.0

Table 1: Recognition results monaural versus binaural presentation; Correct phoneme recognition, deletions, and correct phoneme classification.

3. CONCLUSION

The model presented in this article has to be regarded as an implementation of the rather peripheral stages of a complete model of the human auditory system as depicted in Fig. 1. Principles and knowledge about auditory scene analysis are not yet sufficiently included. Nevertheless the model provides a promising platform for the future development of more sophisticated auditory models which will be able to reproduce auditory streaming and to evaluate properties of hearing events.

Acknowledgments

The work has partly been funded by the BMFT (German Ministery of Research and Technology) under the AuT (Work and Technique) program and the Window-on-Science program.

References

1 Bregman, A, *Auditory Scene Analysis*, MIT PRess, 1990.

2 Bodden, M., “The importance of binaural hearing for noise validation”. In: *Contributions to Psychological Acoustics. Results of the 6th Oldenburg Symposium on Psychological Acoustics*. August Schick (ed.), 1. Ed., Oldenburg: Bibliotheks- und Informationssystem der Carl von Ossietzky Universität Oldenburg, 1993, 537-554.

3 Genuit, K., Blauert, J., "Evaluation of sound environment from the viewpoint of binaural technology", ASJ Symposion, Osaka 1992.

4 Colburn, H.S., Durlach, N.I., "Models of binaural interaction". In: Handbook of Perception, Vol. IV, Hearing, edited by E.C. Carterette and M.P. Friedman. Academic Press, New York, 1978.

5 Blauert, J., *Spatial Hearing - the psychophysics of human sound source localization*, MIT Press, Cambridge, 1983.

6 Stern, R.M., "An overview of models of binaural perception", 1988 National Research Council CHABA Symposium, Washington, D.C., USA, 1988.

7 Blauert, J., “An Introduction to Binaural Technology", in: *Binaural and Spatial Hearing*, R. Gilkey & T. Anderson, Eds., Lawrence Erlbaum, USA-Hilldale NJ,1995 (in press).

8 Zwicker, E., Feldtkeller, R., *Das Ohr als Nachrichtenempfänger*, S. Hirzel Verlag, Stuttgart, 1967.

9 Lindemann, W., "Extension of a binaural cross-correlation model by contralateral inhibition. I. Simulation of lateralization of stationary signals", J. Acoust. Soc. Am. 80, 1986, 1608-1622.

10 Gaik, W., "Combined Evaluation of Interaural Time and Intensity Differences: Psychoacoustical Results and Computer Modeling", J. Acoust. Soc. Am. 94, 1993, 98-110.

11 Pösselt, C., Schröter, J., Opitz, M., Divenyi, P.L., Blauert, J., "Generation of binaural signals for research and home entertainment", Proc. 12th Int. Congr. on Acoustics, Toronto, Vol. 1, 1986, B1-6.

12 Wolf, S., *Untersuchungen zur Lokalisation von Schallquellen in geschlossenen Räumen*, Dissertation, Ruhr-Universität Bochum, 1991.

13 Grabke, J., Blauert, J., "Cocktail-Party-Processors based on Binaural Models", Proc. International Joint Conference on Artificial Intelligence, IJCAI-95, Workshop on Computational Auditory Scene Analysis, 1995, in press.

14 Slatky, H., *Algorithmen zur richtungsselektiven Verarbeitung von Schallsignalen - die Realisierung eines binauralen Cocktail-Party-Prozessor-Systems*, Fortschr.-Ber. VDI Reihe 10 Nr. 286, VDI-Verlag Düsseldorf, 1994.

15 Bodden, M., "Modeling human sound source localization and the Cocktail-Party-Effect", Acta acustica 1(1), 1993, 43-55.

16 Gaik, W., Wolf, S., "Multiple images - psychoacoustical data and model predictions", Proc. of the 8th Int. Symp. on Hearing, edited by H. Duifhuis, J.W. Horst, and H.P. Witt. Academic Press, London, 1988.

17 Hartung, K, “Messung, Verifikation und Analyse von Außenohrübertagungsfunktionen”, Fortschritte der Akustik, DAGA ‘95, DPG-GmbH, Bad Honnef, 1995, in press.

18 Bodden, M, Anderson, T., "A binaural selectivity model for speech recognition", Proc. 4th European Conference on Speech Communication and Technology (Eurospeech), 1995, in press.