Auditory models for spatial impression, envelopment, and localisation

Markus Bodden
Ingenieurbüro Dr. Bodden, Herthastr. 29, D-45131 Essen, Germany

ABSTRACT. In the past years several models which simulate specific abilities of human binaural hearing have been presented in literature. Although most of the models are research models, they can be used for sound source localisation, reproduction of spatial impression, and sound source separation.
Some basic principles of these models will be summarized in this article. Sample applications in the free field, in case of concurrent sources, or in enclosed spaces with reflecting surfaces show the performance - but also the restrictions and limitations of current models.

INTRODUCTION. Research in the human auditory system has led to a variety of auditory models. A group of these models is dedicated to reproduce specific abilities of the binaural aspect of human audition. These models are designed to reveal the spatial distribution of sound sources and to simulate sound source localisation.
A closer look at these models reveals that they are based on common ideas, but offer different implementations, complexities, and features.
The basics of these models will be summarized in the following. However, it is not the intention of this article to give a complete overview of existing models of binaural interaction (for this intention, please refer to literature, e.g. [1], [2], [3]).

BASICS OF BINAURAL AUDITORY MODELS.

Available and used Cues. In principle, different cues are used by the auditory system to localise sound sources or to determine a spatial impression (e.g., [3]):

- Interaul Time Differences (ITD);
- Interaul Intensity Difference (IID);
- Monaural Spectral Cues (e.g., Hebranks and Wright [4], or directional bands as proposed by Blauert [3]);
- Head movements (e.g., Wallach [5]);
- Grouping Cues (Auditory Scene Analysis).

The models presented in literature differ with regard to

- which of these cues they use;
- if more than one cue is used, how they are combined;
- what mechanism they apply to determine and evaluate cues.

Simple models just use one cue (usually ITDs), while more sophisticated ones evaluate combinations or even all of them.
As a result, models show basic principle performance differences:

- models which are pure lateralization models (which means that they are basically restricted to determine the azimuth of a sound source in the frontal horizontal plane - front-back-differentiation and elevation are not evaluated);
- models for complete three-dimensional sound source localisation (azimuth and elevation, not considering distance);
- models which allow localisation based on statistical evaluation, mainly based on temporal averaging of cues. This means that moving sound sources are hard to trace with those models;
- models which are able to reveal the temporal fine structure of sound. They usually are based on a processing in the time domain.

Structures of Models. A basic general structure of binaural models is shown in Fig. 1. Here processing is separated into four different groups.
The first group considers the outer ears. This can be done in several ways:

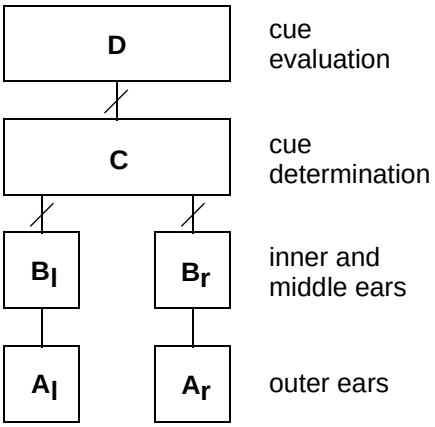

Fig. 1 Basic structure of binaural models.

- for pure simulation models, HRTF catalogues or even simpler simulations of outer ears are used to generate head-related signals from dry mono signals;
- dummy heads can be used to represent a kind of average listener;
- in-ear-microphones can be used to cover the individual HRTFs of a specific listener.

The second block comprises models of the middle and the inner ear. In general simple and deterministic models are used. Probabilistic models simulate the bahaviour of a distribution of hair cells and the movement of the basilar membrane in more detail.
The third block includes the determination of the cues which are used by the model. Most of the models use ITDs, and the most common method to determine them proposed by Jeffress [5] is a crosscorrelation of the signals of the left and the right ear. In general, a corralation can be performed either in the time or in the frequency domain, and thus models can be put into these two different categories.
Although usually a calulation of the ITDs in the frequency domain is computationally more efficient than in the time domain, the calculation in the time domain shows advantages if temporal effects should be considered.
Most of the binaural models use the information delivered by the ITDs, and some of them neglect the information carried by IIDs. This reduction of complexity can be used for simplified models which are designed to reproduce only some basic features of binaural hearing.
Other, more elaborated models also include information carried by IIDs. But, a question which is still a topic of research is the way how interaural information is combined. There are significant hints that the determination of the interaural parameters is performed in seperate units, but that they are evaluated in a common unit (e.g., [3]).
Finally, only some models also consider the influence of monaural cues. These cues can be important with regard to two aspects:

- first, pure monaural sounds (e.g. in headphone representations) are not covered by correlation-based methods;
- second, spectral cues can be important for the elevation perception, as Blauert [3] showed for the directional bands in the median plane.

The last block encounters the evaluation of the cues. In the free-field and for single source situations the cues are not disturbed and can directly be used to analyse the spatial situation. Anyhow, if concurrent sources are present or reflections of the direct sounds at surrounding walls occur, the interaural and monaural cues are also disturbed and do not yield directly to the sound source location. In this case special evaluation methods have to be developed, which separate “desired“ information from “undesired“ information. To do so, grouping cues are sometimes used in algorithms to evaluate the spatial distribution patterns delivered by the models (e.g., [7]).
Since the different models deviate in the way how cues are determined, the evalaution stages show different basic principles. Especially models which perform a calculation in the frequency domain have to get rid of ambiguities of ITDs at higher frequencies. A common way to do so is to integrate information across frequency (e.g., the central spectrum of Raatgever and Bilsen [8], the weighted image of Stern et al [9], the model of Shackleton et al [10], and Duda [11]).
The output of the binaural models show the spatial distribution of sound and thus consist of four dimensions: amplitude, time, frequency, and space.
Some different model types will be shortly described in the following paragraphs.

Time-domain Models. *Deterministic model.* The deterministic model described here was developed by Lindemann [12], Gaik [13], and Bodden [14]. This model is a time domain model which uses ITDs, IIDs and monaural cues. The spatial map produced by the model shows the projection of the 3-dimensional space into the frontal horizontal plane (the azimuth), so that no front-back discrimination is performed.
The principle structure of the model corresponds to Fig. 1. Outer ears can be considered in all forms mentioned in the text above, since the model is able to adapt to individual characteristics of cues of the HRTFs. The middle ear model consists of a simple lowpass filter. The inner ear model is implemented as a bank of bandpass filters with bandwidths corresponding to cri-

tical bands as proposed by Zwicker and Feldtkeller [15], so that 24 frequency bands cover the audio frequency range from 20 Hz to 16 kHz. A lowpass filter with a cutoff frequency of 800 Hz is used to calulate the envelope of signals at high frequencies, and a half-wave rectification and a square root function are applied as a simple hair cell model.
The structure of the binaural kernel is based on the idea of Jeffress [5] performing a crosscorrelation in the time domain. Signals from the left and the right ear move in opposite directions along delay lines and are multiplied at each tap, and a running integration of the products yield the correlation.
The main difference of this model compared to others is the inclusion of a specific inhibition mechanism. Lindemann [12] implemented the so-called contralateral inhibition directly into the crosscorrelation mechanism. Once the signals at a tap contributed to the correlation product they inhibit each other, so that they cannot continue to move along the delay line if they are of equal amplitude. The resulting inhibited crosscorrelation shows advantages compared to a standard crosscorrelation:

- ambiguities of ITDs at higher frequencies are suppressed;
- the width of correlation peaks does no longer depend on frequency - even at very low frequencies sharp correlation peaks can be observed;
- the inhibited crosscorrelation gets sensitive to IIDs.

The latter point is due to the fact that the inhibition is asymmetric if an IID occurs. In this case the stronger signal is not completly inhibited by the weaker signal and continues to move along the delay line, resulting in a shift of the correlation peak to the side.
As a consequence, a special adaptation to head-related signals had to be developed and was presented by Gaik [13]. He introduced an additional individual weighting of the signals on the delay line resulting in the fact that the signals are of equal amplitude at the tap correspondig to their ITD. The weightings are determined in a supervised learning phase.
Bodden extended the model by an evaluation stage to determine the direction of sound incidence and to separate concurrent signals [14]. This stage includes a transformation from the correlation axis to the azimuth, a weighted integration of information across frequency, and some temporal processing. The spatial distribution of sound revealed by the binaural model is used to extract signals from a specific direction of sound incidence out of a mixture of signals. Sample signals are available on a CD [16]. The application of this strategy to automatic speech recognition showed significant improvements of speech recognition rates in adverse conditions [17], and tests with hearing impaired subjects showed that the system is able to increase speech intelligibility in concurrent-speaker situations [18].

Probabilistic Model. A model which was specially designed for localisation in rooms was presented by Wolf [19]. In this coincidence model based on the evaluation of ITDs, only the information carried by rising slopes in the signals is used. The rising slopes are transformed into ideal peaks by some specific processing, and the correlation mechanism is reduced to a coincidence detection.
In doing so, information carried by reflections is suppressed, so that the direction of sound incidence (here also the azimuth in the frontal horizontal plane) can be determined. Nevertheless, this model requires temporal averaging, and there is no continuous flow of information as in the model described above. The output patterns of the model can thus not directly be used to separate concurrent sources.

Statistical Model. Slatky [20] presented a special model to determine the direction of sound incidence for two concurrent signals. He investigated the behaviour of ITDs for two narrowband concurrent signals (as critical band signals are) and found that a relation exists between the mean and standard deviation of parameters extracted from the analytical signal and the directions of sound incidence. Similar to the approach explained for the deterministic model he also used this information to separate the concurrent sources, achieving about the same performance for two signals. But, since this mathematically oriented model is based on the assumption of two sources, the performance degrades for situations with more sources.

Frequency-domain model. A simple and computationally efficient frequency domain model for broadband signals was presented by Kunz and Bodden [21]. Like in other models, frames of the signals are transformed by means of a discrete fourier transform into the frequency domain, interaural parameters are extracted and compared to a database. Since the basic assumption of the model is that signals are rather broadband, the interaural differnences are only evaluated at specific frequencies. The model is able to determine the direction of sound incidence (including elevation and front-back discrimination) in anechoic environment, but the performance descreases fast if concurrent sources or reverberation are present.

SAMPLE APPLICATIONS OF BINAURAL AUDITORY MODELS. Most of the binaural models presented in literature have the status of research models. They are more intended to help explaining abili-

ties of the human auditory system than for technical applications.
Usually these models are based on structures which have been derived from psychoacoustic and physiological experiments in the free field. As a result, models have mostly been tested under free-field conditions, but less in enclosed or even small spaces.
Applications presented here have been performend with a model corresponding to the deterministic model described above (if not indicated otherwise).

Free Field. Binaural models offer optimal performance in the free field. The reason is obvious: the cues used by the models are not affected by the environment.
In case of a single sound source the cues are even "ideal". This means that the cues correspond perfectly to the data used to train or adapt the model. Thus - if the cues itself are not ambiguous - perfect performance can be expected.
An example for a pattern produced by the model is depicted in Fig. 2. It shows the excitation in one critical band (300 - 400 Hz) as a function of time. The correlation axis of the model was replaced by an azimuth axis using a transformation presented by Bodden [7]. The source was broadband noise at an azimuth of 30^{o}.

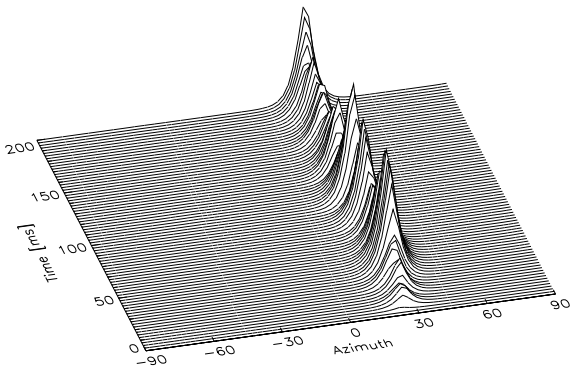

Fig. 2 Binaural excitation pattern in one critical band (300 - 400 Hz).

It can be seen that the azimuth is clearly determined, that correlation peaks are sharp even in this low frequency band, and that the temporal envelope of the signal is maintained in the amplitude of the excitation peaks.
If multiple sound sources are present, the performance of models usually decreases. The most important factor influencing the performance is the short-term correlation of the concurrent sources - the higher the correlation is, the worse is the performance. Fig. 3 shows an example for two concurrent signals (two speakers) at about the same average signal level. In this figure the predicted direction of sound incidence is depicted. The prediction is based on a weighted integration of the information provided by each critical band. It can be seen that both directions are correctly determined.

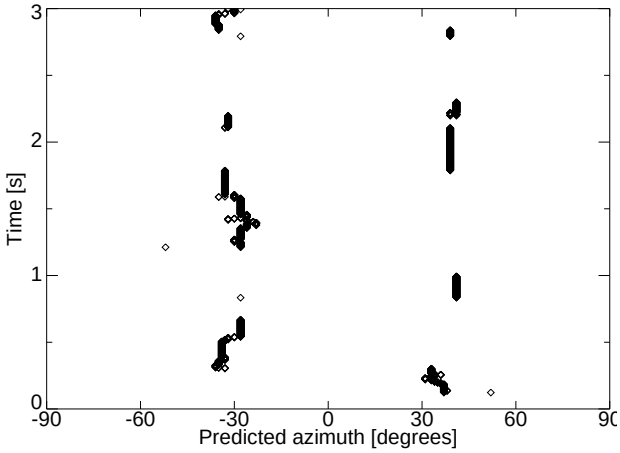

Fig. 3 Predicted direction of sound incidence as a function of time. Signal: two concurrent speakers at 40^{o} and -30^{o}.

Reflections. From a physical point of view reflections can be regarded as interference if the localisation of a sound source should be investigated. But, in contrast to the interference caused by different sound sources reflections are highly correlated with the direct sound. It is one of the most sophisticated abilities of the human auditory system to deal with interference and especially reflections. The way how the human auditory system selects exactly the information which is necessary for localisation, but also uses the information of reflections, e.g. for the spatial impression, is not yet known in detail. Although some aspects of the precedence effect are quantified yet, we still have a lack of understanding the complete underlying processing, so that a complete model could not yet be implemented. As a result, the localisation performance of models breaks down earlier than localisation abilities of humans.
Fig. 4 shows an example for the predicted direction of sound incidence in a reverberant room (reverberation time 0.9 s). In this case a speaker was positioned at an azimuth of about 50^{o}close to a reflecting wall.
It can be seen that the model is not able to completely suppress the reflections from the wall - the predicted azimuth switches between the position of the sound source and the direction of incidence corresponding to the dominant wall reflection.

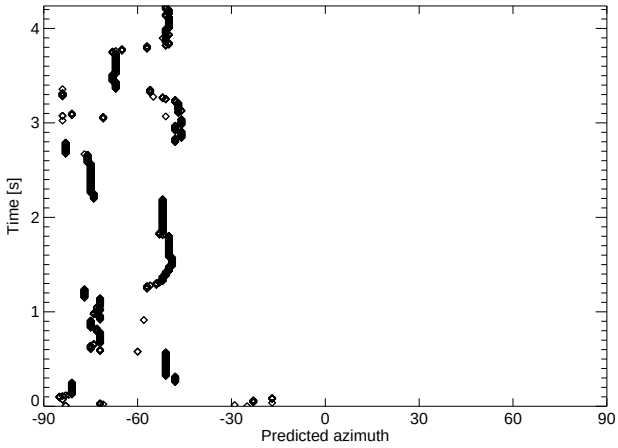

Fig. 4 Predicted direction of sound incidence as a function of time. Signal: speaker at 50^{o} close to a wall in a small room.

Room Acoustics. As stated above, reverberant conditions reduce the performance of the model. But, on the other hand, the excitation patterns produced by the model can be used to visualise the temporal and spatial distribution of reflections in rooms.
In this context Blauert, Lehnert and Bodden [22] showed the feasibility of a binaural model for room acoustics planning. They used binaural room impulse responses of different concert halls and compared the excitation patterns produced by the binaural model. Fig. 5 shows examples of corresponding spatial maps produced by the model. The figure depicts the excitation as a function of azimuth and time in critical band no 5 (400 - 510 Hz). The input signals were binaural impulse responses of a lecture hall (taken from [23]) which were simulated with a binaural room simulation model. In the top graph, the absorption coefficients of all walls were set to 30%, while in the bottom one they were set to 70%. The figure displays the spatial distribution for one critical band, and corresponding graphs can be printed for each other frequency band, so that the frequency-dependency of reflections can also be revealed.
Since the input signal was a binaural room impulse response the graphs visualize the spatial distribution of room reflections as a function of time and frequency. The influence of the wall absorption coefficients can clearly be seen when comparing the two graphs.
The output patterns of the model can thus also be used to evaluate the quality of room acoustics.

Small Spaces. In case that the surfaces of small spaces have high absorption coefficients localisation can be reproduces by the models. Fig. 6 shows an example for the spatial map produced by the simple frequency domain model presented by Kunz and Bodden [21]. The source (a male speaker) was positioned at 30^{o}azimuth and 0^{o}elevation in a small soundproofed chamber (about 2x2.5x2 m). The excitation presented in the figure was averaged over about 1 s. It can be seen that the position of the sound source is correctly identified, but that also other receptive fields show some excitation. If the number of sources is not known, it would thus be difficult to state how many sources are active in that situation.
In cases that walls have no high absorption coefficients, localisation becomes a difficult task. Then reflections are arriving at the eardrums with a short time delay to the direct sound and a high amplitude. The time delay can even be so short that the precedence effect (e.g., [3]) is no longer valid, and that even the human listener is no longer able to determine the direction of sound incidence.
In contrast to the simulation presented in Fig. 4, in small spaces, e.g., in the interior of a car, reflections are arriving from various directions of incidence at the same time. In this case the determination of the direction of sound incidence becomes a hard task. Due to

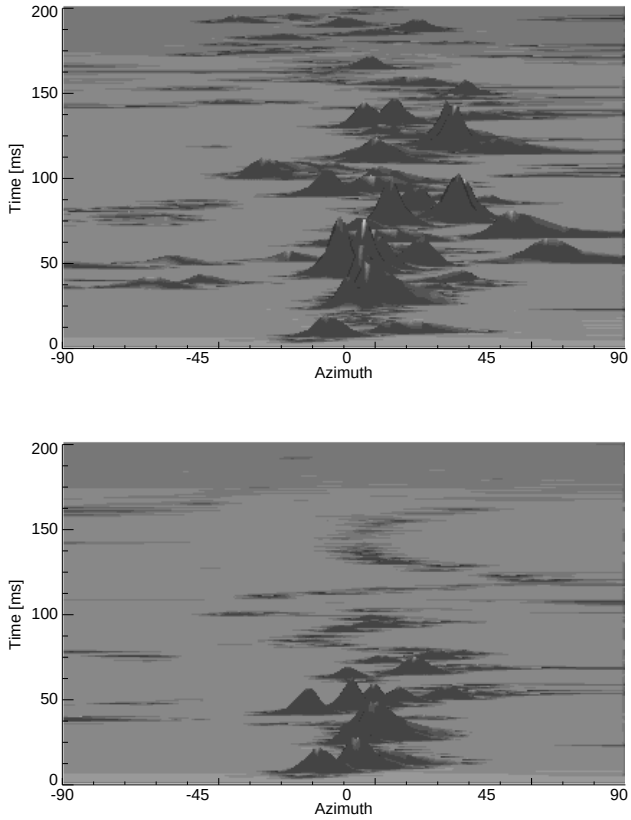

Fig. 5 Spatial distribution of sound, critical band no. 5 (400-510 Hz). Signal: binaural impulse responses of a simulated lecture hall (taken from [23]).
Top: surfaces with an absorption coefficient of 30%
Bottom: surfaces with an absorption coefficient of 70%

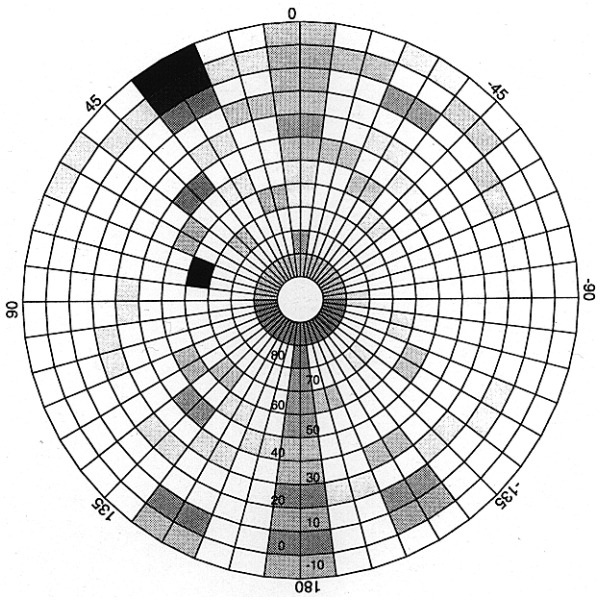

Fig. 6 Spatial map produced by a frequency domain model for a source of 30^{o} azimuth and 0^{o} elevation in a small room (2x2.5x2 m). The intensity of color codes the excitation in each reception field.

strong early reflections (e.g., from the window close to the drivers ear) and standing waves in the compartment interaural cues are heaviliy influenced is such spaces.

It had been stated before that sophisticated binaural models evalute the combination of ITDs and IIDs and compare them to learned values. The models will thus not be able to determine the direction of sound incidence correctly, but they are likely to predict the broadening of hearing events and the increasing diffuseness of spatial impression.

An adaptation of the model to the specific interaural cues inside of a car would be possible from a theoretical point of view, but it has to be considered that they depend strongly of the exact position of the microphones. Even a small change of the positions can change the IIDs dramatically (e.g., due to standing waves), so that this would not make sense in practise.

A general drawback of the models which hinders a performance similar to human localisation is that head movements can not be utilized. It is known since long time (e.g., Wallach [5]) that these head movements are important for human sound source localisation. The listener moves his head in a controlled manner, and the resulting changes in the interaural and monaural cues allow to resolve ambiguities and to determine the direction of sound incidence.

SUMMARY AND OUTLOOK. The performance of models of binaural hearing presented in literature shows that a substancial progress has been achieved in the past years. Anyhow, the performance of the models especially in small spaces with strong early reflections still has to be improved. The kernel question in this context is how the auditory systems manages to differentiate between useful and interfering information.

In order to answer this question, current activities review the physiological origin of binaural processing in more detail. By means of applying virtual sound sources to electrophysiological experiments with guinea pigs Hartung and Sterbing [24] are able to investiagte spatial tuning in the inferrior colliculus even for complex but well controlled sound fields, and questions like front-back confusions, the role of monaural cues, and the precedence effect can be addressed on a neural level. The results of these experiments yield strategies to improve the struture of binaural auditory models [25].

In small spaces the localisation performance of current models is rather poor. But, as shown above, the models can well be used to evaluate the spatial distribution of reflections. This information can for example be used to evaluate the acoustics in rooms and to derive strategies to improve the acoustics.

Acknowledgements. Most of the work presented in this article has been performed during the authors stay at the “Institut für Kommunikationsakustik“ of the Ruhr-University Bochum, Germany, headed by Prof. Blauert.

Literature. [1]Colburn, H.S., Durlach, N.I. (1978). Models of binaural interaction. In: Handbook of Perception, Vol. IV, Hearing, edited by E.C. Carterette and M.P. Friedman. Academic Press, New York.

[2] Stern, R.M. (1988). An overview of models of binaural perception. 1988 National Research Council CHABA Symposium, Washington, D.C., USA.

[3] Blauert, J. (1997). Spatial Hearing - the psychophysics of human sound source localization. Revised edition, MIT Press, Cambridge.

[4] Hebrank, J., Wright, D. (1974). Spectral cues used in the localization of sound sources in the horizontal plane”, Journal of the Acoustical Society of America 56, 1829-1834.

[5] Wallach, H. (1940): The role of head movements and the vestibular and visual cues in sound localization. Journal of Experimental Psychology, 27, 339-368.

[6] Jeffress, L.A. (1948). A place theory of sound localization. J. Comp. Physiol. Psych. 61, 468-486.

[7] Bodden, M. (1995): Binaural Modeling and Auditory Scene Analysis. IEEE Signal Processing Society, 1995 Workshop on Appl. of Signal Processing to Audio and Acoustics, Mohonk Mountain House, New Paltz, NY.

[8] Raatgever, J., Bilsen, F.A. (1986): A central spectrum theory of binaural processing. Evidence from dichotic pitch. J. Acoust. Soc. Am. 80, 428-441.

[9] Stern, R.M.; Zeiberg, A.S.; Trahiotis, C. (1988): Lateralization of complex binaural stimuli A weighted-image model. J. Acoust. Soc. Am. 84, 156-165.

[10] Shackleton, T.M.; Meddis, R.; Hewitt, M.J. (1992): Across frequency integration in a model of lateralization. J. Acoust. Soc. Am. 91, 2276-2279.

[11] Duda, R.O. (1997): Estimating azimuth and elevation from the interaural head related transfer-function. In: Binaural and spatial hearing, R.H Gilkey and T. Anderson (eds.), Erlbaum, N.J.

[12] Lindemann, W. (1986). Extension of a binaural cross-correlation model by contralateral inhibition. I. Simulation of lateralization of stationary signals. Journal of the Acoustical Society of America 80, 1608-1622.

[13] Gaik, W. (1993). Combined Evaluation of Interaural Time and Intensity Differences: Psychoacoustical Results and Computer Modeling. Journal of the Acoustical Society of America 94, 98-110.

[14] Bodden, M. (1993): Modeling Human Sound Source Localization and the Cocktail-Party-Effect. Acta Acustica 1(1), 43-55.

[15] Zwicker, E., Feldtkeller, R. (1967). Das Ohr als Nachrichtenempfänger, S. Hirzel Verlag, Stuttgart.

[16] Bodden, M. (1996): Auditory Demonstrations of a Cocktail-Party-Processor. Acustica united with acta acustica Vol. 82 no 2, 356-357.

[17] Bodden, M., Rateitschek, K. (1996): Noise-robust speech recognition based on a binaural auditory model. Proc. ESCA workshop on the “Auditory Basis of Speech Perception”, Keele, UK, July 1996.

[18] Bodden, M. (1997): Binaural hearing and hearing impairment: relations, problems, and proposals for solutions. Seminars in Hearing Vol. 18 No. 4, 375-391.

[19] Wolf, S. (1991): Untersuchungen zur Lokalisation von Schallquellen in geschlossenen Räumen, Dissertation, Ruhr-Universität Bochum.

[20] Slatky, H. (1992): Binaurale Signalverarbeitung bei Anwesenheit mehrerer Schallquellen: Untersuchungen zum Cocktail-Party-Prozessor-Problem", Dissertation, Ruhr-Universität Bochum.

[21] Kunz, O.; Bodden, M. (1996): Ein rechenzeiteffizientes Modell zur Lokalisation von Schallquellen in Realzeit. Fortschritte der Akustik - DAGA’96, DPG-GmbH, Bad Honnef, 364-365.

[22] Blauert, J.; Bodden, M.; Lehnert, H. (1992): Binaural Signal Processing & Room Acoustics Planning. IEICE Trans. Fundamentals. Vol. E75-A, No. 11, Nov. 1992, 1454-1459.

[23] The AUDIS catalogue of human HRTFs. CD-Rom, Documenta Acustica (09/DE2), D-Herzogenrath, 1998 (see http://eaa.essex.ac.uk/eaa).

[24] Hartung, K.; Sterbing, S.J. (1997): Generation of virtual sound sources for the electrophysiological characterization of auditory spatial tuning in the guinea pig. Acoustical Signal Processing in the Central Auditory System, Plenum Press, N.Y., 407-412.

[25] Hartung, K.; Sterbing, S.J. (1998): Ein neurophysiologisch motiviertes Modell zur Lokalisation von Schallquellen. Fortschritte der Akustik, DAGA ’98. DPG-GmbH, Bad Honnef, in press.