subito copyright regulations

Copies of articles ordered through subito and utilized by the users are subject to copyright regulations. By registering with subito, the user commits to observing these regulations, most notably that the copies are for personal use only and not to be disclosed to third parties. They may not be used for resale, reprinting, systematic distribution, emailing, web hosting, including institutional repositories/archives or for any other commercial purpose without the permission of the publisher.

Should delivery be made by e-mail or FTP the copy may only be printed once, and the file must be permanently deleted afterwards.

The copy has to bear a watermark featuring a copyright notice. The watermark applied by subito e.V. must not be removed.

Modeling human sound-source localization and the cocktail-party-effect

Markus Bodden

Lehrstuhl für allgemeine Elektrotechnik & Akustik, Ruhr-Universität Bochum, D-44780 Bochum, Germany

(Received 5 March 1993; accepted 4 May 1993)

Abstract. — A person with normal hearing is able to understand speech signals even under acoustically adverse conditions, e.g., if the desired speech signals interfere with signals of concurrent speakers. This property of human hearing, which is often refered to as the "Cocktail-Party-Effect", is partly made possible by binaural processing in the auditory system. In this article an algorithm is described which allows to simulate the Cocktail-Party-Effect electronically – based on our current knowledge of the human auditory system and on earlier modeling work (Lindemann 1986, Gaik 1990, 1993). The algorithm produces simulations of neural excitation patterns including the spatial distribution of neural excitation. Further analysis is then rendered by a model of more central stages of the signal processing in the auditory system. As a result it is possible to predict the azimuth of sound incidence with respect to the listener's head. Further, the algorithm estimates parameters of the incoming signals specified by their respective azimuths. These parameters are then used to control the transfer function of a time-variant optimum filter (Wiener-Filter), such as to enhance one desired signal out of the spatial distribution of concurrent signals. The performance of the algorithm is demonstrated in terms of its directional selectivity. Results of two different comprehensibility tests with hearing-impaired subjects show considerable improvement of comprehensibility. The algorithm is expected to be of advantage for many technological applications, e.g., automatic speech recognition, advanced hearing-aids and hands-free telephony.

Pacs numbers: 43.66Qp — 43.60Lq — 43.72Ew

1. Introduction

The performance of the human binaural system with regard to speech perception is indeed striking. Speech signals can be understood even if they are subject to heavy interference by noise, reverberation or concurrent speech. This property of human hearing is often refered to as the "Cocktail-Party-Effect". The need for a technological replication of these binaural signal-processing capabilities for many application is obvious, as the following examples may demonstrate.

- Typical application scenarios for speech recognizers include noisy environments in which current recognition algorithms tend to show a significant decrease in performance.
- Hearing-impaired persons must come to terms with reduced intelligibility in noisy environments. Conventional hearing-aids do not offer much help with respect to this task.
- In hands-free telephony, a desired speaker is to be picked out of number of source signals whilst avoiding audible feed-back in the transmission chain.

Various methods have been proposed to tackle the problem of enhancing the intelligibility of speech interfered by noise. For an overview and comparison see, e.g., Lim (1983). In general, the algorithms can be divided into the following categories.

- Single-channel methods to suppress noises with stationary spectra, e.g., spectral subtraction methods.
- Adaptive noise-cancelling algorithms, most of which need an additional channel for dealing with the noise to be suppressed.
- Microphone-array systems – often combined with adaptive algorithms.

These methods have in common that they are based on purely algorithmic, i.e. signal-theory approaches, more or less disregarding our knowledge of the human auditory system. In general their range of application is limited to restricted, well-defined sound-field characteristics – e.g., stationary noise sources – or to unpractical large arrangements of microphone arrays (for an exception, see Soede, 1990). In particular, poor results are achieved in cases where the interfering signal is concurrent speech.

An ideal speech enhancement method should be able to suppress any kind of interfering noise, much as the human auditory system itself, which can thus be regarded as a reference. Consequently, the method described in this article simulates to a large extent human binaural signal processing, thus avoiding any general restrictions with regard to sound-field or signal characteristics.

Besides simulating the Cocktail-Party-Effect, the system proposed here is able to predict sound-source azimuths. This is of interest for further research in hu-

man sound localization and for technical applications besides speech enhancement. For example, a novel binaural noise-measurement technique employing psychoacoustical properties of the human auditory system is in the process of being developed [1] (Genuit 1991; Notbohm et al. 1992; Remmers and Prante 1991; Bodden and Gaik 1991).

The Cocktail-Party-Processor as proposed here performs an analysis of the spatial distribution of sounds. It can be regarded as an extension of a prior model of sound localization as developed by Lindemann (1986) and Gaik (1990, 1993). This prior model is rather complex in terms of signal processing, thus only a brief overview can be given here. Please refer to the literature for details.

2. The concept of the cocktail-party-processor

Figure 1 depicts the general concept of the Cocktail-Party-Processor (see also Bodden 1990, 1992). As compared to the model of Lindemann (1986) and Gaik (1990, 1993), a module has been added which models more central stages of human binaural signal processing. Amongst other things, this new module performs analyses of neural excitation patterns with the aim of predicting sound-source azimuths. Further, a Wiener-Filter is introduced as a signal processing tool to produce the estimate of a desired signal out of a mixture of concurrent signals. In the following paragraphs the system and its modules are described in more detail.

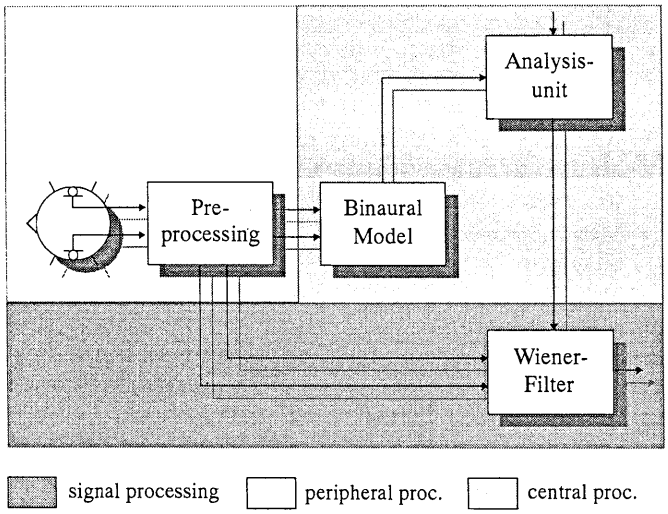

Figure 1. The concept of the Cocktail-Party-Processor.

[1] project funded by the BMFT (Federal Ministery of Research and Technology) under the AuT program (Work and Technique): "Entwicklung einer Meßtechnik mit Berücksichtigung der psychoakustischen Eigenschaften des Nachrichtenempfängers "Menschliches Gehör" zur physiologischen Bewertung von Lärmwirkung" (Development of a measurement technique for the physiological valuation of the effects of noise taking into account psychoacoustical properties of the human auditory system).

3. The binaural model

Overviews on models of binaural perception can be found in Colburn & Durlach (1978), Blauert (1983), and Stern (1988). The binaural model employed for the processing proposed in this article has been developed by Lindemann (1986) and Gaik (1990, 1993). It is based on an interaural cross-correlation function that has been extended by powerful processing algorithms like contralateral inhibition and an adaptation to head-related transfer functions. Due to these extensions the model is able to reproduce both, influences of interaural differences in time (IDT) and interaural intensity differences (IID) and, what is more important, of the combination of both interaural parameters. Gaik and Wolf (1988) compared the results of simulations computed with this model to data derived from psychoacoustical experiments. Their investigation showed that the model is able to predict fused hearing events as well as multiple events that can occur in artificial listening situations. The modules simulating the auditory periphery and the binaural processor will be described in the following paragraphs.

3.1. *The structure of the binaural model*

The structure of a binaural model as proposed by Blauert (1983) is depicted in Figure 2. It consists of modules simulating processing of outer, middle and inner ear, followed by a binaural processor and a pattern recognizer modeling higher stages of the human auditory system.

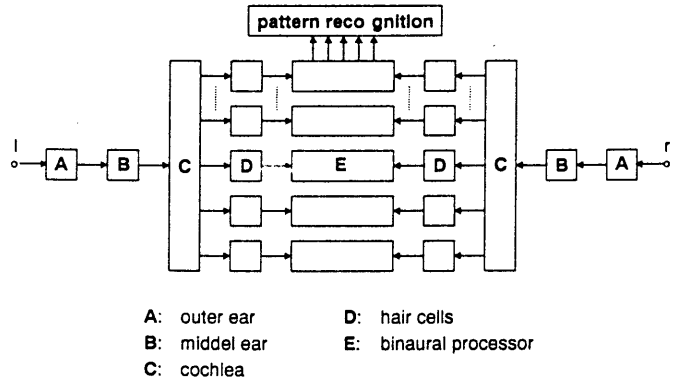

Figure 2. The structure of the binaural model proposed by Blauert (1983).

3.1.1. *Outer and middle ear*

Processing of the outer ears can be described in form of directional dependent transfer functions. These transfer functions are called head-related transfer functions (HTF's) and introduce interaural differences in time (IDT) and intensity (IID) into the signals that can be recorded at the eardrums of a listener. Those differences are fundamental cues evaluated by the binaural auditory system. The head-related transfer functions show interindividual differences that are considered by the binaural processor as will be described in 3.1.4.

For some simulations presented in this paper ear signals have been synthesized. Pösselt et al. (1986) measured catalogues of head-related transfer functions covering 122 directions of incidence from the upper hemisphere. He either positioned subminiature microphones in the earcanals or used the microphones of the dummy heads to measure HTF's of human subjects or artificial heads, respectively. Binaural signals can then be produced by convolution of signals recorded in an anechoic chamber with the head-related transfer functions corresponding to the desired direction of incidence. For a psychophysical evaluation of this technique see Wenzel (1992).

The transfer function of the middle ear does not depend on the direction of incidence. Measurements of Letens (1989) showed that it can be represented by a linear lowpass filter.

Table 1. *Cutoff-frequencies of the filters according to the critical bands after Zwicker and Feldtkeller (1967).*

critical band no.	f_u [Hz]	f_o [Hz]
1	20	100
2	100	200
3	200	300
4	300	400
5	400	510
6	510	630
7	630	770
8	770	920
9	920	1080
10	1080	1270
11	1270	1480
12	1480	1720
13	1720	2000
14	2000	2320
15	2320	2700
16	2700	3150
17	3150	3700
18	3700	4400
19	4400	5300
20	5300	6400
21	6400	7700
22	7700	9500
23	9500	12000
24	12000	15500

3.1.2. *Inner ear*

Processing of the inner ear can be modeled in a quite detailed manner using comprehensive models of the basilar membrane movement (Michel, 1988). For the signal processing purposes described in this article an accurate reproduction of the processing of the peripheral system is not required. Satisfactory results can be achieved by employing a simplified model. As a rough approximation of the cochlear processing, we use a set of $N = 24$ bandpass filters. The bandwidths of the filters correspond to critical bands as proposed by Zwicker and Feldtkeller (1967). The cutoff frequencies of the unsymmetrical filters that cover the frequency range from 20 Hz to 15.5 kHz (4th order slope to low frequencies, 10th order to high frequencies, after Duifhuis, 1972) are shown in table 1. The transfer functions of the digital filters are depicted in Figure 3.

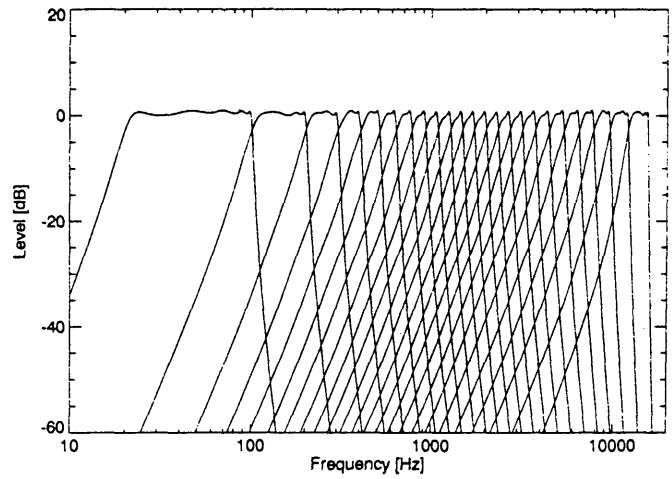

Figure 3. Transfer functions of the filters used as basilar membrane model. 24 unsymmetrical digital Chebychev IIR-filters of 4th and 10th order according to the critical bands of Zwicker and Feldtkeller (1967) are used.

In each critical band, time signals are calculated as rough approximations of the firing probabilities of nerve fibres. First, the bandpass signals are half-wave rectified. To take into account that envelopes of signals are processed at high frequencies, the signals are then fed into lowpass filters with cutoff frequencies of 800 Hz. Saturation of the firing probabilities of nerve fibres are included by extracting the square root of the obtained time functions.

3.1.3. *The Binaural Processor*

Following an idea of Jeffress (1948) the binaural processor developed by Lindemann (1986) is based on a running interaural cross-correlation of the signals produced by the inner ear models of the left and right ear. The processing scheme is shown in Figure 4. The cross-correlation mechanism is realized in parallel for the 24 critical bands in the time domain. The signals from the left and right ear move in opposite directions along two delay lines. The contents of the delay lines are multiplied at each tap and a running integration is performed to produce the cross-correlation function. Lindemann introduced a powerful extension by applying a mechanism known from neural processing: inhibition. Transfered to the correlation algorithm he called this mechanism contralateral inhibition. It 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 output of the binaural processor, the running inhibited interaural cross-correlation function, can be regarded as a simulation of neural excitation. We will call the output neural excitation from now on.

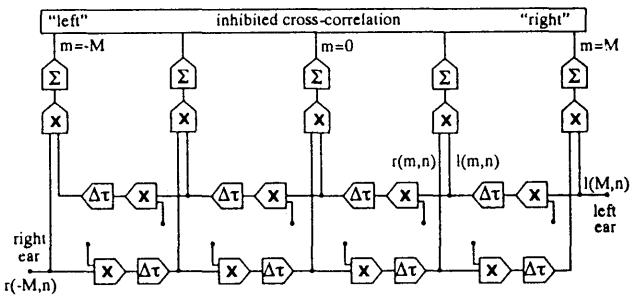

Figure 4. Calculation of the running interaural cross-correlation function with contralateral inhibition as proposed by Lindemann (1986). n describes the time index, m the tap number on the delay line and M the length of the delay line in points (usually $M = 40$ for a sampling rate of 40 kHz). r and l are the signals supplied from the preprocessing as described in 3.1.2.

The contralateral inhibition offers a variety of advantages:

- the periodicity of the resulting cross-correlation function is suppressed and therefore ambiguities are avoided. A weighting of the correlation function that emphasizes the more central peaks, as proposed by Stern et al. (1988) and performed by Shackleton et al. (1992), can be omitted. In contrast to their models, single peaks occur for single sound sources (except for unnatural conditions as described later);
- a contrast enhancement of the output patterns can be observed;
- the resulting correlation function becomes sensitive to interaural intensity differences.

The latter point is due to the fact that contralateral inhibition becomes unsymmetrical if an interaural intensity difference occurs. Therefore, both interaural differences in time and intensity are analyzed in one processing step.

A further important extension of the correlation mechanism is the integration of monaural processors sited at the end of the delay lines. Therefore, hearing events in unnatural listening conditions can be predicted, for example pure monaural events and multiple events that can occur in headphone representation.

3.1.4. *Adaptation to head-related transfer functions*

Gaik (1990, 1993) introduced an important extension to the binaural processor, the adaptation to head-related transfer function. If head-related signals are fed into the Lindemann model, unfused hearing events may be predicted. This effect is due to the strong dependency of the interaural intensity difference on frequency. In the Lindemann model an interaural intensity difference leads to unsymmetrical contralateral inhibition. This results in a shift of the correlation peak to the side or even a splitting up into two peaks. The idea of Gaik was that the processor has to adapt to the individual interaural parameters of the used set of head-related transfer functions. He calculated interaural time and intensity differences of the head-related transfer function in each critical band for 122 directions of incidence from the upper hemisphere. The data show that the interaural parameters form natural combinations, which means that the interaural intensity difference can be calculated from the interaural time differences. For frequencies higher than 800 Hz the IID does not increase monotonically with increasing azimuth, but decreased for azimuths bigger than 60 degrees. Here the relation becomes ambiguous.

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. Using this method, fused hearing events are predicted from model simulations using head-related signals. Furthermore, Gaik was able to show that even if unnatural combinations are used, simulation results agree with the performance of the human auditory system. In psychoacoustical experiments with contradictory interaural parameters, e.g., trading experiments, multiple events can occur. Simulations calculated with the binaural model show the same qualitative results (Gaik and Wolf, 1988). In the present stage of development the model is not able to resolve source locations in the cone of confusion.

The way how the auditory system combines the interaural parameters IDT and IID is still not exactly evaluated. Data recently published by Wightman and Kistler (1992) show that IDT cues dominate IID cues in the low frequency region (as reported by others before; see Blauert, 1983) and that for frequencies higher than 5 kHz IID cues dominate IDT cues. A similar weighting of the importance of the interaural cues is performed by the model due to contralateral inhibition. IDT cues are nearly independent on frequency. For low frequencies IID cues are quite small and IDT cues dominate the position of the peaks on the correlation axis. As IID cues increase with frequency they have an increasing effect on the strength of the contralateral inhibition and dominate IDT cues at high frequencies.

4. Modeling higher stages of the human auditory system

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 a pattern recognition algorithm

as depicted in Figure 1 is able to:

1) predict the projection of the positions of hearing events into the horizontal plane. This leads to a model of sound source localization;
2) discriminate between neural excitations due to sound sources from different directions of incidence. This leads to a Cocktail-Party-Processor.

A system to predict sound source localization and a Cocktail-Party-Processor based on the binaural model are described in the following paragraphs.

4.1. *Localization*

The proposed algorithm does not model the processing of the human auditory system in detail. It is designed to reproduce the ability of the auditory system to attach a position in space to a hearing event. Here, the description of this position is restricted to its projection into the frontal horizontal plane, that is, the azimuth.

In case of one sound source in a non-reverberant environments its direction of incidence can directly be determined from the interaural delay corresponding to the peak positions of the neural excitation patterns. Figure 5 shows an example of a neural excitation pattern for this sound field situation. Each curve represents the neural excitation of one critical band, the abscissa shows the interaudal delay.

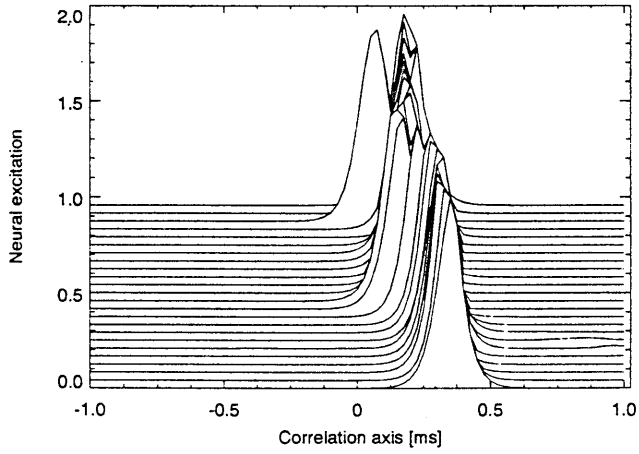

Figure 5. Neural excitation patterns for one sound source in non-reverberant environment. The abscissa shows the interaural delay, each curve represents the excitation of one critical band. The band with the lowest centre frequency is plotted at the bottom. The sound source azimuth was 30 degrees.

If more than one sound source or reflections are present, the neural excitation patterns are more difficult to evaluate. Due to the interference, peaks occur at various and quickly changing positions. Therefore, a pattern recognition algorithm has to extract information that is relevant for the localization process. This task is performed by the analysis unit depicted in Figure 1. It includes the following processing steps that are described in the following paragraphs:

- a correlation-azimuth transformation replaces the correlation axis representing the interaural delay with an axis representing the azimuth;
- a running average 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;
- a determination of azimuths.

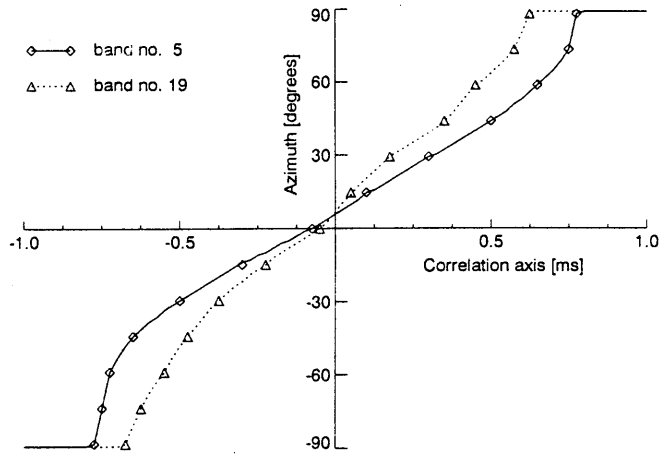

Figure 6. Correlation-azimuth transformation in critical band no. 5 (solid line) and 19 (dashed line). Marked points represent values derived in the learning phase of the model, curves show the interpolated transformation rule.

4.1.1. *Correlation-azimuth transformation*

The processing step introduced here transforms the correlation axis to an axis that directly represents the azimuth. In order to calculate the transformation function, the relation between interaural delays and corresponding azimuths has to be determined. This is done in a supervised learning phase, where simulations of the binaural model are used to set up a transformation rule describing this relation. For this learning phase we use white noises from directions of incidence from the frontal horizontal plane, e.g., signals with azimuths ranging from $-90°$ to $+90°$ in $15°$ steps. They are fed into the model and the interaural delays corresponding to the peak positions in each critical band are determined from the resulting neural excitation patterns. Then, linear interpolation is used to calculate the azimuth as a function of the interaural delay. Figure 6 shows two examples of transformation functions, here for critical band no. 5 and 19. The transformation functions for the other critical bands look quite similar.

4.1.2. *Running average*

The system proposed here should not be restricted to fixed sound sources but also be able to trace moving sound sources. Attempts to derive time constants employed by the human auditory system in dynamic spatial resolution tasks led to various results. In a recent publication Chandler and Grantham (1992) measured

minimum audible movement angles and argue that time constants of up to 500 ms for broadband sources and even up to 1500 ms for narrowband sources are necessary for an "optimal performance of the binaural system". Other experimental data suggest time constants of about 100 ms (Grantham and Wightman, 1978; Grantham, 1982; Blauert, 1983). For the simulations presented in this paper a time constant of 100 ms was used.

4.1.3. *Combination of critical bands*

We define a total neural excitation as the weighted sum of the neural excitations produced in each critical band. The weighting can be interpreted as a spectral weighting of neural excitations. A similar weighting has been introduced by Raatgever (1980) and Raatgever and Bilsen (1986) and was used by Stern et al. (1988). They derived a spectral weighting function from results of localization experiments reported by Raatgever (1980). These data show best localization accuracy in the spectral region centered around 600 Hz. As measurements of Raatgever cover a frequency range up to 1200 Hz, no weighting can be derived for higher frequencies. In addition, his data have been derived from experiments using pure interaural differences in time; interaural intensity differences have not been considered. The spectral weighting applied here is derived in a different manner: results of model simulations itself are used. Similar to the calculation of the transformation function presented in 4.1.1., the weighting of each critical band is determined in a learning phase. Signals produced as described in 4.1.1. are fed into the model and the correlation-azimuth transformation is performed. At high frequencies the relation between IDT and IID can become ambiguous (see 3.1.4.), and therefore the predicted azimuths may differ from the sound source azimuths. The predicted azimuths $\tilde{\varphi}$ are directly represented by the peak positions of neural excitations on the azimuth scale. We can now calculate the differences between sound source azimuths φ and predicted azimuths $\tilde{\varphi}$ for the $D = 13$ directions of incidence (azimuths ranging from $-90°$ to $+90°$ in $15°-$ steps) and define a mean error $\bar{A}_i$ in each critical band i :

$$\bar{A}_i = \frac{1}{D}\sum_{d=1}^{D} |\varphi_d - \tilde{\varphi}_d| \qquad (1)$$

The mean error $\bar{A}_i$ describes the reliability of information provided by each critical band. We decided to define a weighting W_i of each critical band that exponentially depends on the mean error. The shape of the weighting function can be controlled by the parameter A_τ :

$$W_i = \mathrm{e}^{\frac{\bar{A}_i}{A_\tau}} \qquad (2)$$

Figure 7 shows an example of the weighting function for a value of $A_\tau = 15°$.

The total neural excitation $e_T(\varphi, t)$ is calculated as the weighted sum of the neural excitations $e_i(\varphi, t)$:

$$e_T(\varphi, t) = \sum_{i=1}^{N} e_i(\varphi, t).\ W_i, \qquad (3)$$

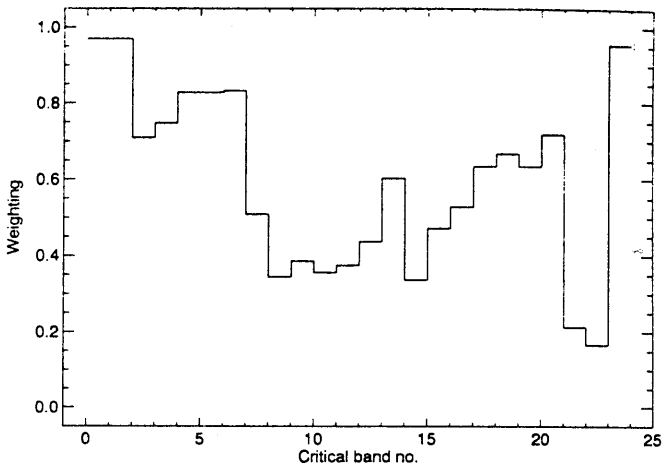

Figure 7 Spectral weighting of the neural excitation used to calculate the total neural excitation for one catalogue of head-related transfer functions.

with time t and azimuth φ

4.1.4. *Determination of azimuths*

Each peak observed in the total neural excitation represents a candidate for an azimuth of a hearing event. In order to decide whether a candidate does represent a hearing event we check the time course of the amplitude of the total neural excitation. An ascending amplitude can be observed if new directional information appear. If a candidate for a new azimuth is proposed, the derivative of amplitude is calculated. If the derivative exceeds a threshold the candidate is accepted.

4.2. *The Cocktail-Party-Processor*

In a Cocktail-Party-Situation the signals of a desired sound source and one or more concurrent sound sources interfere. The input signal $x(t)$ to the Cocktail-Party-Processor can be expressed as the sum of the desired signal $s(t)$ and the interference $n(t)$:

$$x(t) = s(t) + n(t). \qquad (4)$$

In order to build an estimate $\hat{s}(t)$ of the signal of the desired sound source the time varying transfer function $H(f,t)$ of a Wiener-Filter is estimated. The spatial distribution of neural excitation enables the processor to discriminate between excitation due to the desired sound source and excitation due to interfering sound sources.

Under the assumption that $s(t)$ and $n(t)$ are orthogonal the transfer function $H(f,t)$ of a time varying Wiener-Filter can be determined by calculating the ratio

$$H(f,t) = \frac{C_{ss}(f,t)}{C_{xx}(f,t)}, \qquad (5)$$

where $C_{ss}(f,t)$ and $C_{xx}(f,t)$ represent the power density spectra of the desired signal $s(t)$ and the distorted signal $x(t)$, respectively. Since the calculations are performed in parallel in 24 critical bands the transfer function $H(f,t)$ can be split up into 24 transfer functions $H_i(f,t)$:

$$H_i(f,t) = \frac{C_{ss,\,i}(f,t)}{C_{xx,\,i}(f,t)}. \tag{6}$$

In order to determine these transfer functions the power density spectra $C_{ss,\,i}(f,t)$ and $C_{xx,\,i}(f,t)$ have to be estimated. For bandwidths corresponding to critical bands we are not able to calculate estimates that depend on frequency. In this case the estimate of the power density spectrum, $\hat{C}_{ss,\,i}(f,t)$, can be expressed by the root-mean-square (rms) value $s_{\text{eff},i}(t)$ of the bandfiltered signal $s_i(t)$:

$$\hat{C}_{ss,\,i}(f,t) = s^2_{\text{eff},i}(t). \tag{7}$$

Thus we have to estimate rms-values from the neural excitation patterns. In order to do this we define the total binaural excitation $E_{x,i}(t)$ as the sum of the neural excitation $e_i(\varphi,t)$ on the azimuth axis:

$$E_{x,\,i}(t) = \sum_{j=-90°}^{90°} e_i(j,t). \tag{8}$$

The desired binaural excitation $E_{s,\,i}(t)$ is determined by windowing the neural excitation with a window Θ centered at the azimuth φ_s corresponding to the direction of incidence of the desired sound source and summing up the resulting excitation:

$$E_{s,\,i}(t) = \sum_{j=-90°}^{90°} e_i(j,t).\Theta_i(\varphi_s) \tag{9}$$

The window function $\Theta_i(\varphi_s)$ can be desired from neural excitation patterns of a model simulation of one sound source in non-reverberant environment. The shape of the peaks resulting from this simulation can be regarded as ideal peak shapes and serve as the window function $\Theta_i(\varphi_s = 0)$. The window function $\Theta_i(\varphi_s)$ is determined by centring $\Theta_i(\varphi_s = 0)$ around the azimuth φ_s.

The determination of the binaural excitations $E_{x,\,i}(t)$ and $E_{s,\,i}(t)$ is depicted in Figure 8. The figure shows a schematic example for one critical band.

As the neural excitation is based on a cross-correlation function, the rms value of $s_i(t)$ can be estimated from the binaural excitation $E_{s,\,i}(t)$:

$$\hat{s}^2_{\text{eff},\,i}(t) = E_{s,\,i}(t), \tag{10}$$

and the rms value of $x_i(t)$ from the binaural excitation $E_{x,\,i}(t)$:

$$x^2_{\text{eff},\,i}(t) = E_{x,\,i}(t). \tag{11}$$

Combining (6), (7), (10) and (11) the transfer function $\hat{H}_i(t)$ representing an estimate of the Wiener-Filter $H_i(f,t)$ can be calculated as

$$\hat{H}_i(t) = \frac{E_{s,\,i}(t)}{E_{x,\,i}(t)} \tag{12}$$

As $\hat{H}_i(t)$ does not depend on frequency it becomes a time varying weighting factor that we call $g_i(t)$ to avoid confusion:

$$\hat{H}_i(t) = g_i(t) \tag{13}$$

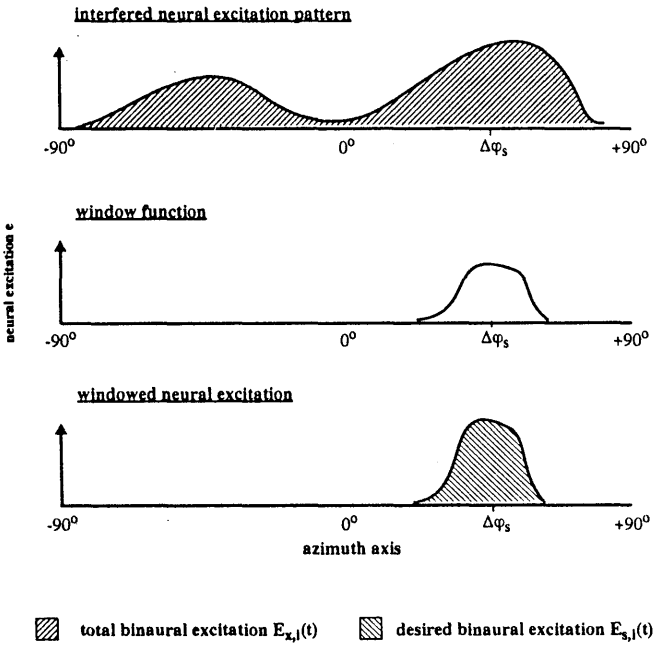

Figure 8 Symbolic presentation of the determination of the weighting factors from neural excitations. Upper curve: neural excitation in one critical band. The azimuth of the desired sound source is marked by $\Delta\varphi_s$. Middle curve: Weighting function centred at the position of the azimuth of the desired sound source. Lower curve: Estimate of the neural excitation corresponding to the desired sound source. The hatched areas represent the corresponding binaural excitations.

The calculation of the weighting factors $g_i(t)$ is valid if $s(t)$ and $n(t)$ are orthogonal. This assumption seems reasonable since the calculation is based on short intervals of the signals (1 to 10 ms). However, correlation between $s(t)$ and $n(t)$ may arise, and in this case the performance of the separation algorithm will decrease.

The output signal $\hat{s}(t)$ of the Cocktail-Party-Processor is built as the sum of the weighted critical band signals derived from the preprocessing filterbank. 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:

$$\hat{s}_{l,\,r}(t) = \sum_{i=1}^{N} x_{l,\,r,\,i}(t) \cdot g_i(t) \tag{14}$$

The indices l and r describe the left and right channel of the binaural input signal.

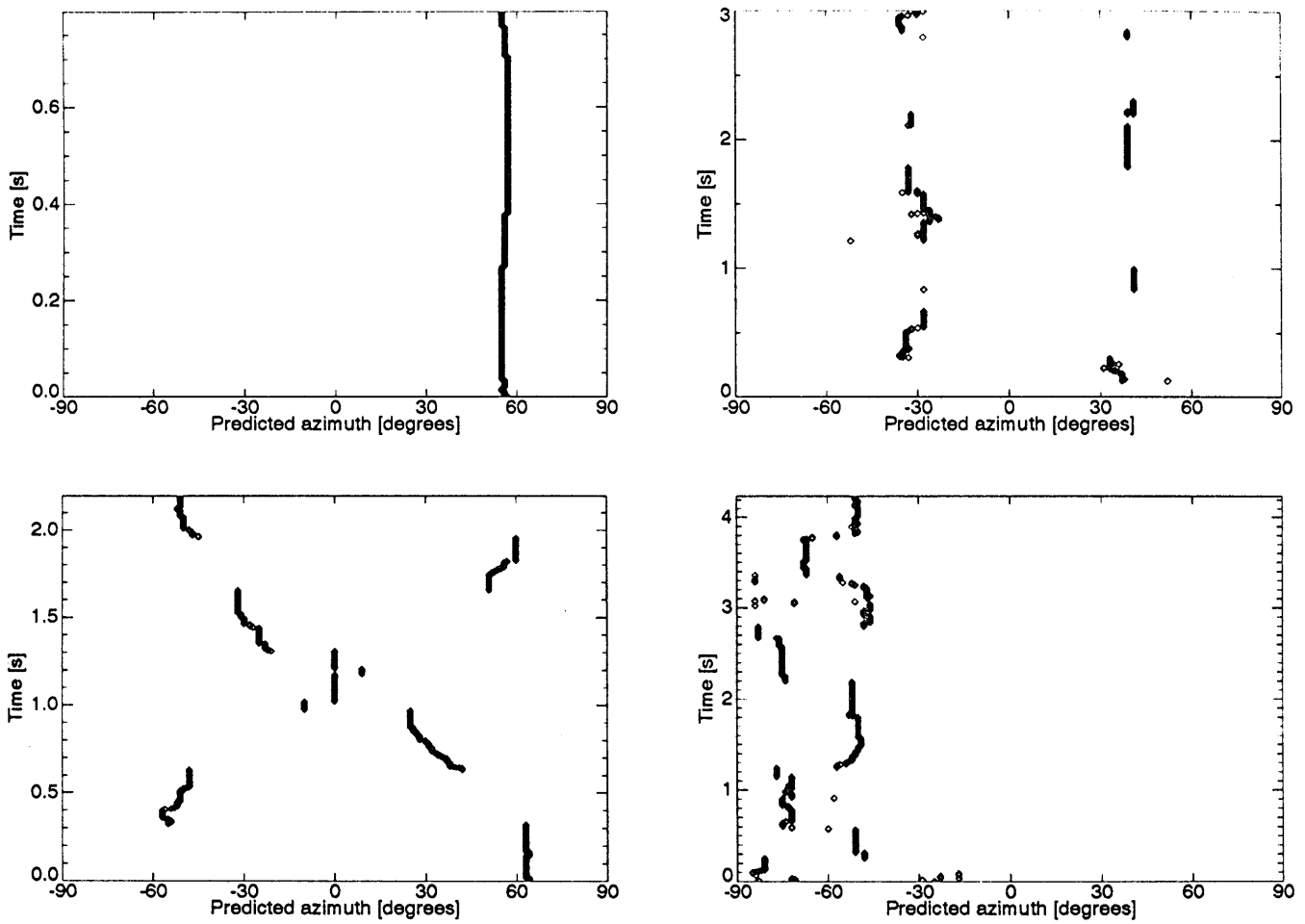

Figure 9. Predicted localization for different conditions. The ordinate shows time from the start of the analysis. Upper left: dummy head recording of one sound source (white noise) positioned at an azimuth of 60° in non-reverberant environment. Upper right: simulation of two simultaneous speakers at azimuths of −30° and +45°. Lower left: simulation of two alternating sound sources (uncorrelated white noises) moving in opposite directions. Lower right: two microphone recording of one sound source in reverberant environment at an azimuth of −45°. (See text for a further description).

For the simulations presented in this paper the desired sound source azimuth has been user-defined. A combination with the analysis system presented in 4.1. would enable the system to find the directions of incidence of sound sources automatically, so that even moving sound sources could be treated.

5. Results

The performance of the localization model will be evaluated by comparing real sound source azimuths and predicted azimuths for different sound field situations. To judge on the performance of the Cocktail-Party-Processor beam forming and results of comprehensibility tests will be presented.

5.1. *Localization*

Figure 9 shows four examples of the output of the system to predict localization. The time constant for the moving average involved in the analysis algorithm was set to 100 ms for all examples. The diagrams show the predicted azimuth on the abscissa and the time from the start of the analysis on the ordinate:

1) The upper left diagram shows the results for a dummy head recording of one sound source (white noise) in non-reverberant environment. The sound source was fixed at an azimuth of 60°. The predicted azimuth differs only up to 5° from the sound source azimuth and is provided immediately after starting the analysis.

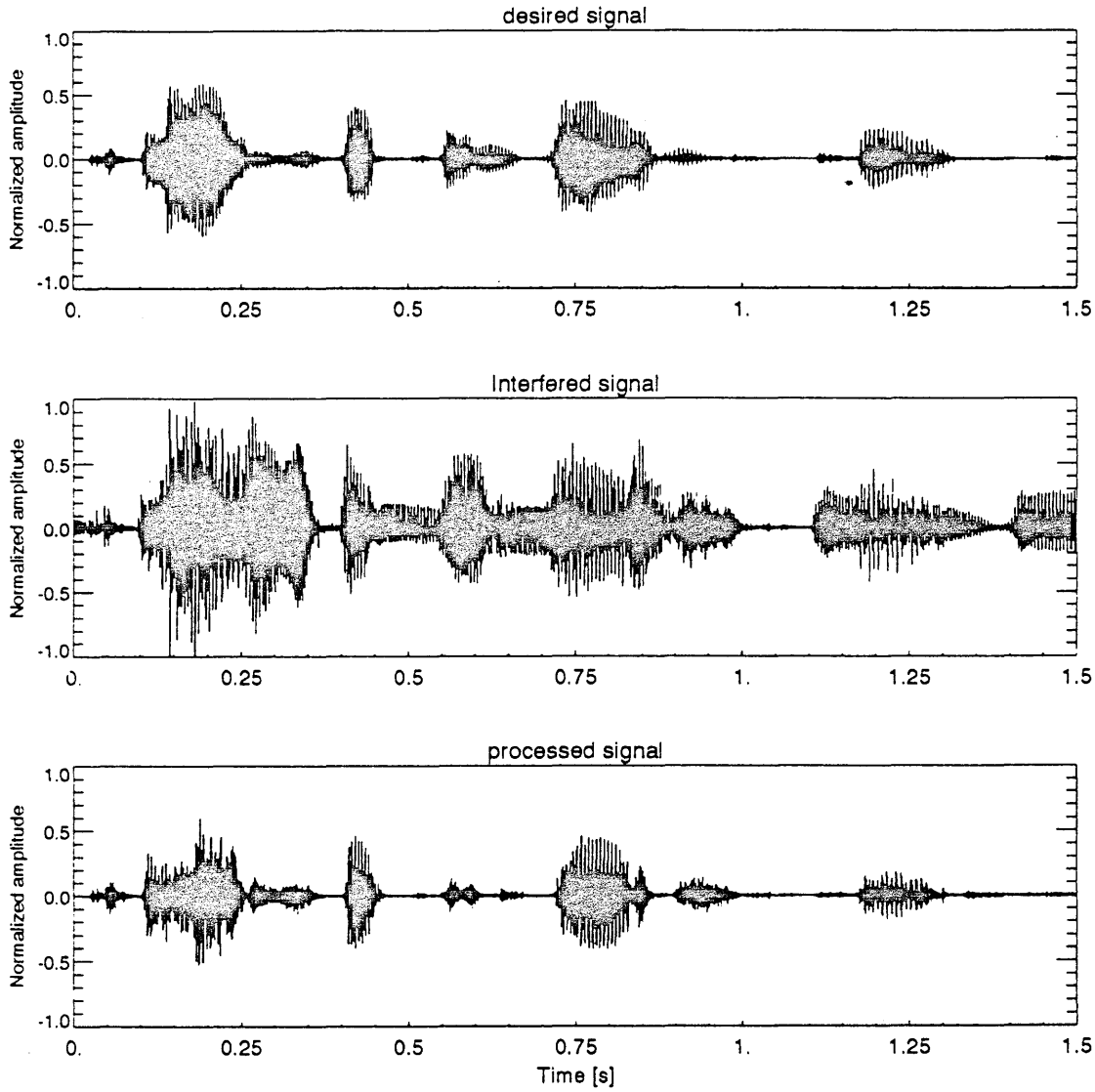

Figure 10. Upper curve: speech signal of the desired speaker (azimuth 0°, right channel of a dummy head recording). Middle curve: The same signal interfered by a concurrent speech signal (azimuth 30°, right channel). Lower curve: output of the Cocktail-Party-Processor.

2) The upper right diagram shows the results for two simultaneous speech signals at a mean signal-to-noise ration of 0 dB. The signals have been synthesized by convolution of speech signals of a male and a female speaker with head-related transfer functions corresponding to the azimuths −30° and +45°. The analysis results show that the predicted azimuths correspond to the real speaker positions. Since the analysis system predicts only one azimuth for one moment in time, the predicted azimuth sometimes 'switches' between the left and right position. Incorrect localization information is nearly perfectly suppressed, which means that nearly no azimuths are predicted that do not correspond to real speaker positions.
3) The lower left diagram shows the results for a simulation of two alternating moving sound sources. The signals have been produced with a binaural mixing console. Such a device calculates head-related signals by convolution of input signals with head-related transfer functions and mixes several of them together to produce a binaural output signal. For this simulation, two uncorrelated noised have been used. The noise samples had a length of 350 ms, followed by a pause of 300 ms. The two signals were alternating, but they overlapped for short periods of 50 ms. They have been treated by the binaural mixing console in such a manner that one signal moves from the right side to the left and the second in opposite direction from the left side to the right. It can be seen from the simulation results that the movement of both sources is traced by the algorithm.

4) The lower right diagram shows the results for a two microphone recording [2] of one sound source (a male speaker) in reverberant environment. The sound source position was at −45° close to a reflecting wall in a square room with a reverberation time of about 1 s. The sound source position is identified, but a second hearing event closer to the left is predicted. This second predicted hearing event may be due to strong reflections at the wall that are not recognized as reflections of the speech sound source.

The results prove the ability of the system to predict localization of sound sources in non-reverberant environments even for multiple and moving sound sources. In reverberant environments the model may predict additional hearing events. Here, further processing steps have to be introduced to refine the performance of the system. Research on the precedence effect (Clifton et al., 1984; Clifton, 1987; Blauert et al., 1989; Wolf, 1991) has proven that onsets of signals are important for sound source localization in reverberant environments. Further investigations have to be carried out on this topic.

5.2. *Cocktail-Party-Processing*

An example of the suppression of interfering speech signals is shown in Figure 10. A speech signal of a male speaker has been convolved with head-related transfer functions corresponding to an azimuth of 0°. The signal is shown in the upper curve. An interfering speech signal of a female speaker (but with similar pitch) has been convolved with head-related transfer functions corresponding to an azimuth of 30° and binaurally added to the desired signal at a mean signal-to-noise-ratio of 0 dB. The signal is plotted in the middle curve. The output signal of the Cocktail-Party-Processor is shown in the lower curve. All curves depict the right channel.

The interfering signal is nearly absolutely suppressed and only slight changes in the course of the processed signal compared to the desired speech signal can be observed.

5.2.1. *Beam Forming*

The ability of the Cocktail-Party-Processor to suppress interfering sound sources can be expressed by means of beam forming. In this context beam forming is defined as the attenuation of a signal as a function of its azimuth of incidence when the frontal direction (azimuth $\varphi = 0°$) is defined as the desired direction of incidence. Figure 11 shows the resulting beam forming in different critical bands: the solid line corresponds to band no. 2 (100-200 Hz), the broken line to band no. 6 (510-630 Hz), dashes and points to band no. 13 (1.72-2 kHz) and the dotted line to band no. 18 (3.7-4.4 kHz). The input signals for the Cocktail-Party-Processor have been calculated by convolution of white noise with head-related transfer functions of the frontal horizontal plane. The

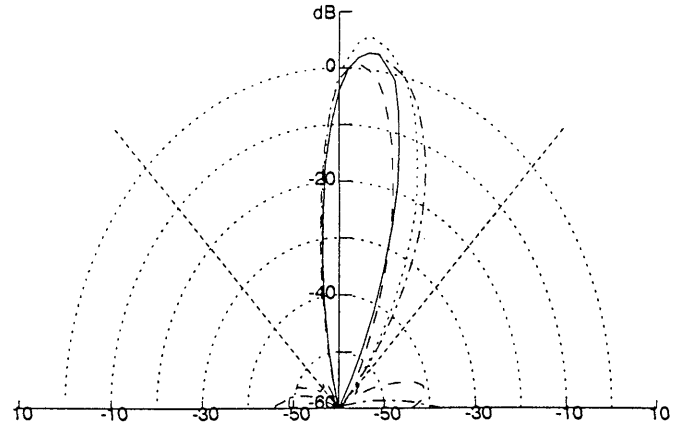

Figure 11. Beam forming of the Cocktail-Party-Processor. The solid line corresponds to band no. 2 (100-200 Hz), the broken line to band no. 6 (510-630 Hz) dashes and points to band no. 13 (1.72-2 kHz) and the dotted line to band no. 18 (3.7−4.4 kHz). The frontal direction (0°) has been defined as the desired direction.

beam forming curves have been determined by interpolation of the attenuations derived from the output signals of the processor when the frontal direction was user-defined as the desired direction. The resulting beams are narrow even in the first critical band and the width of beams is nearly independent on frequency. The slight shift to the right side that can be observed in all beams is due to asymmetries in the head-related transfer functions used for signal generation.

5.2.2. *Comprehensibility Tests*

In attempts to apply conventional speech enhancement algorithms, e.g., comb filtering methods, to the Cocktail-Party-problem deteriorations of intelligibility have been reported (Lim, 1983) even for improved signal-to-noise ratios. Therefore, intelligibility tests had to be performed. The Cocktail-Party-Processor presented here may be of interest for the future development of advanced hearing-aids. Hence we decided to perform the tests with hearing impaired subjects.

Some standardized tests methods to measure the intelligibility of interfered speech use interfering signals with speech-like characteristics. Some methods use noises with the long term spectra of speech or signals that have been produced by the overlay of different speech signals after eliminating the pauses (Kollmeier, 1991). For the test of the Cocktail-Party-Processor presented here signals of competitive speakers have been chosen as interfering signals. Recent investigations on speech assessment have shown that the use of meaningless speech material offers some advantages (Jekosch, 1990). Using this kind of material the influence of interindividual deductive capabilities of subjects listening to test stimuli can be excluded: if subjects identify only a part of the stimuli, they are not able to derive the remaining part using their language knowledge as it may be possible for meaningful speech stimuli. Hence, this test is directed towards com-

[2] Two-microphone recordings can be processed by the binaural model, too. As those recordings include no interchannel intensity differences the adaptation to the interindividual parameters described in 3.1.4 can be omitted.

prehensibility and describes the percentage of correctly identified stimuli.

A first test was developed using semantically unpredictable sentences (Benoît et al. 1989). These sentences consist of meaningful words in grammatically correct order forming sentences without predictable meanings. Two situations were considered at a mean signal-to-noise ratio of 0 dB:

– 3 speakers: −45°, female; 0°, female (desired speaker); 30°, male

– 2 speakers: −30°, male; 30°, female (desired speaker)

A total number of 50 sentences were recorded in an anechoic chamber and adjusted to equal root-mean-square (rms) level. Binaural signals were produced by convolving them with head-related transfer functions of the corresponding directions of incidence. The interfered signals were presented dichotically, whereas the processed signals were presented diotically. The subjects used their own hearing-aids in all tests, and open HD-414 headphones were placed over the hearing-aids to present the signals. At the beginning of the tests the subjects were asked to adjust a confortable sound level while keeping the usual setting of the hearing-aids. In each test condition 10 sentences were presented in random order. Four untrained hearing impaired subjects participated in all tests, and no feedback was given and the different test conditions were presented in mixed order to avoid learning effects. The subjects showed symmetrical or slightly asymmetrical hearing losses of 60 dB to 70 dB and were supplied with 2 hearing-aids. The task of the subjects was to write down on a sheet of paper what they comprehended. Then, the correctly identified works have been counted. The averaged results of the test are depicted in Figure 12. The bars represent comprehensibilities before (light) and after (dark) processing for the 2-speaker situation on left and the 3-speaker situation on the right. The gain in comprehensibility due to processing is significant and even higher in the 3-speaker situation.

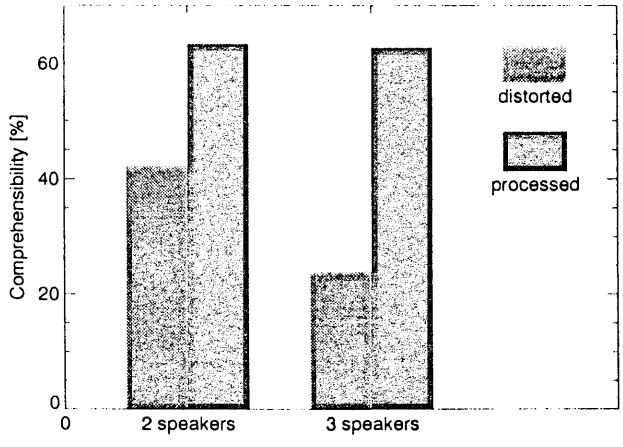

Figure 12. Result of the sentence test: mean comprehensibilities for four hearing-impaired subjects before (light barks) and after (dark bars) processing.

An additional test has been developed using 100 meaningless works (Jekosch, 1990). The words are build as consonant-vowel-consonant (CVC) cluster combinations using clusters based on statistics of the German language. This test offers the advantage that confusions between different clusters can be extracted and the results can be analyzed in a more detailed manner. The signals have been recorded in an anechoic chamber using two male speakers. The interfered signals have been produced by convolving the signals with head-related transfer functions corresponding to +/−30° and adding them up in random order at a mean signal-to-noise ratio of 0 dB. Each speaker served 50 times as desired speaker. The interfered signals were dichotically presented via headphones, the processed signals were diotically presented using a different random order. Again, the subjects task was to write down what they comprehended and the correctly identified words and clusters were counted. The averaged results of the comprehensibility tests with five hearing impaired subjects are depicted in Figure 13. Light bars represent comprehensibilities before and dark bars after processing. The results are presented for the words on the left and for the different clusters on the right.

A significant gain in comprehensibility ranging from 12 to 25% can be observed. It can be remarked that comprehensibilities for the consonant clusters are enhanced in the same range as for the vowel clusters.

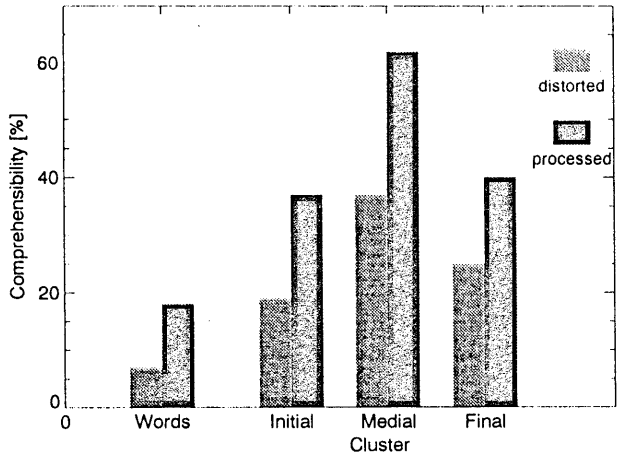

Figure 13. Result of the word test: mean comprehensibilities for five-hearing impaired subjects before (light barks) and after (dark bars) processing. Left: words. Right: Results for the initial, medial and final clusters.

6. Conclusion

Based on a comprehensive binaural model simulations of higher stages of the human auditory system have been proposed. Evaluating the neural excitation patterns produced by the binaural model some important aspects of human sound source localization and the Cocktail-Party-Effect can be reproduced.

The analysis of the neural excitation patterns involving strategies that are supposed to take part in the human auditory system leads to predictions of azimuths of

hearing events. The predicted azimuths are in good agreement with the sound source azimuths even for multiple and moving sound sources. In reverberant environments additional hearing events due to strong reflections may be predicted. To overcome this problem it will be necessary to further understand the processing underlying the precedence effect.

The spatial distribution of neural excitation enables the system to distinguish between excitations due to sound sources from different directions of incidence. Using this information the proposed Cocktail-Party-Processor estimates the time-varying transfer function of an optimal filter. The resulting processing algorithm performs narrow beam forming covering the whole audio frequency range from 20 Hz to 15.5 kHz. In speech tests with hearing impaired subjects a significant increase in comprehensibility was found.

Acknowledgements.

The author wishes to thank Prof. Dr.-Ing. J. Blauert for the support of this study and Dr. Georges Canévet and the reviewers for their helpful comments on the manuscript. The work has partly been funded by the BMFT (German Ministery of Research and Technology) under the AuT (Work and Technique) program.

References

BENOÎT CHR., VAN ERP A., GRICE M., HAZAN V., JEKOSCH U., (1989) "Multilingual Synthesizer Assessment using Semantically Unpredictable Sentences", Proc. of Eurospeech '89, Paris, Vol. 2, 633-636.

BLAUERT J. (1983) "Spatial Hearing", MIT-Press, Cambridge, Mass.

BLAUERT J., CANÉVET G., VOINIER TH, (1989) "The precedence effect: No evidence for an "active" release process found", J. Acoust. Soc. Am. 85, 2581-2586.

BODDEN M. (1990) "A Concept for a Cocktail-Party-Processor", Proc. ICSLP '90, Kobe, Japan, 285-289.

BODDEN M. (1992) "Binaurale Signalverarbeitung: Modellierung der Richtungserkennung und des Cocktail-Party-Effektes", Fortschr.-Ber. VDI Reihe 17 Nr. 85. VDI Verlag, Düsseldorf.

BODDEN M. (1992) "Cocktail-Party-Processing Concept and results", Proc. of the 14th. ICA, Beijing, China, L3-2.

BODDEN M., GAIK W., (1991) "Ein binaurales Modell zur Verarbeitung kopfbezogener Signale", Fortschritte der Akustik, DAGA '91, Bochum. DPG-GmbH,Bad Honnef, 785-788.

CHANDLER D.W., GRANTHAM D.W., (1992) "Minimum audible movement angle in the horizontal plane as a function of stimulus frequency and bandwidth, source azimuth, and velocity", J. Acoust. Soc. Am. 91, 1624-1636.

CLIFTON R.K., MORRONGIELLO B.A., DOWD J.M., (1984) "A developmental look at an auditory illusion; The precedence effect", Dev. Psychobiol. 17, 519-536.

CLIFTON R.K. (1987) "Breakdown of echo suppression in the precedence effect", J. Acoust. Soc. Am. 82, 1834-1835.

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.

DUIFHUIS H. (1972) "Perceptual Analysis of sound", Dissertation Technische Hoogeschool Eindhoven, The Netherlands.

GAIK W. (1990) "Untersuchungen zue binauralen Verarbeitung kopfbezogener Signale", Fortschr.-Ber. VDI Reihe 17 Nr. 63. Düsseldorf VDI-Verlag.

GAIK W. (1993) "Combined Evaluation of Interaural Time and Intensity Differences: Psychoacoustical Results and Computer Modeling", accepted for publication in J. Acoust. Soc. Am.

GAIK W., WOLF S. (1988) "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.

GENUIT K. (1991) "Gehörgerechte Lärmbewertung", Fortschritte der Akustik, DAGA '91, Bochum. DPG-GmbH, Bad Honnef, 75-92.

GRANTHAM D.W., WIGHTMAN F.L., (1978) "Detectability of varying interaural temporal differences", J. Acoust. Soc. Am. 63, 511-523.

GRANTHAM D.W. (1982) "Detectability of time-varying interaural correlation in narrow-band noise stimuli", J. Acoust. Soc. Am. 72, 1178-1184.

JEFFRESS L.A. (1948) "A place theory of sound localization", J. Comp. Physiol. Psych. 61, 468-486.

JEKOSCH U. (1990) "A weighted intelligibility measure for speech assessment", Proc. ICSLP '90, Kobe, Japan, 973-976.

LETENS U. (1989) "Über die Interpretation von Impedanzmessungen im Gehörgang anhand von Mittelohrmodellen", Dissertation Ruhr-Universität Bochum.

LIM J.S. (1983) "Speech Enhancement", Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

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

MICHEL D. (1988) "A model for peripheral auditory preprocessing", in: Mechanics of hearing, ed. by J.P. Wilson and D.T. Kemp.

NOTBOHM G., SCHWARZE S., JANSEN G. (1992) "Noise evaluation based on binaural hearing", Proc. of the 14th. ICA, Beijing, China, H2-2.

PÖSSELT C., SCHRÖTER J., OPITZ M., DIVENYI P.L., BLAUERT J., (1986) "Generation of binaural signals for research and home entertainment", Proc. 12th Int. Congr. on Acoustics, Toronto, Vol. 1, B1-6.

RAATGEVER J. (1980) "On the binaural processing of stimuli with different interaural phase relations", Doctoral dissertation, Technische Hogeschool Delft, The Netherlands.

RAATGEVER J., BILSEN F.A. (1986): A central spectrum theory of binaural processing. Evidence from dichotic pitch, J. Acoust. Soc. Am. 80, 428-441.

REMMERS H., PRANTE H., (1991) "Untersuchung zur Richtungsabhängigkeit der Lautstärkeempfindung von breitbandigen Signalen", Fortschritte der Akustik, DAGA '91, Bochum, DPG-GmbH, Bad Honnef, 537-540.

SHACKLETON T.M., MEDDIS R., HEWITT M.J., (1992) "Across frequency integration in a model of lateralization", J. Acoust. Soc. Am. 91, 2276-2279.

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

SOEDE W. (1990) Improvement of speech intelligibility in noise: development and evaluation of a new directional hearing instrument based on array technology. Doctoral dissertation, TU Delft.

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

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.

WENZEL E.M. (1992): Localization in Virtual Acoustic Displays. Presence 1 (1), Teleoperators and Virtual Environments, 80-107.

WIGHTMAN F.L., KISTLER D.J. (1992) "The dominant role of low-frequency interaural time differences in sound localization", J. Acoust. Soc. Am. 91, 1648-1661.

WOLF S. (1991) "Untersuchungen zur Lokalisation von Schallquellen in geschlossenen Räumen", Dissertation, Ruhr-Universität Bochum.

ZWICKER E., FELDTKELLER R. (1967) "Das Ohr als Nachrichtenempfänger", S. Hirzel Verlag, Stuttgart.