Improved Binaural Model for Localization of Multiple Sources

Magnus Schäfer, Mohammad Bahram, Peter Vary

Institute of Communication Systems and Data Processing (ind), RWTH Aachen University
Email: {schaefer,bahram,vary}@ind.rwth-aachen.de
Web: www.ind.rwth-aachen.de

Abstract

An improved binaural hearing model is proposed that consists of a physiologically motivated signal processing step and a subsequent cognitive model. The model is shown to be capable of correctly detecting and tracking sources while blindly determining the number of active sources based on temporal and spectral information.

The complete model is of interest in all areas that need to consider the capabilities of the human hearing system while the ability to determine the number of active sources makes it a logical enhancement for source separation and spatial signal processing such as adaptive beamforming.

1 Introduction

Spatial hearing is a research topic that has received continuous interest from different scientific areas throughout the last century with Rayleigh's duplex theory [1] as an important first milestone highlighting the role of interaural time difference (ITD) and interaural level difference (ILD). Throughout the 20th century, many research efforts concentrated on descriptive evaluations of the properties and capabilities of the human hearing system, both for monaural and binaural perception – overviews on many of the experiments that were carried out can be found in [2, 3].

The next major step towards accurately modeling the way that humans perceive sounds in general and spatial properties in particular took place after also considering the increasing knowledge about human physiology. An overview on the way that acoustic events are processed in the human auditory system and many of the derived models can be found in [4].

Most of the binaural perception models can be grouped according to two fundamental theories:

- The coincidence-based model of Jeffress [5]
- The equalization-cancellation model of Durlach [6]

Most modern models of binaural hearing are based on Jeffress's introduction of special neural units called *coincidence cells*. These record coincidences in neural firings from hair cells from both ears within one frequency band. Furthermore, the neural signals are delayed by a small amount that is fixed for a given fiber pair. In other words, the ITD of a single stimulus will be coded by means of a so-called *internal time difference* τ. This specific τ refers to the coincidence cell having the highest response activity (highest fire rate). By finding the most common internal time difference over a certain frequency range, the direction of the source can be estimated. Jeffress's model is actually a coarse representation of the structure of the auditory nervous system, which was unknown at that time and could only be detected physiologically significantly later [7].

The equalization-cancellation model was originally designed for modeling binaural masking level differences. It was later shown that it is a feasible way to explain various other auditory phenomena as well. The capabilities and limitations of this model are discussed in detail in [8].

In the remainder of this paper, the coincidence-based model and its extensions by Blauert and his colleagues [9] will be improved even further by incorporating both a weighting function for different sub-bands as well as a distribution function of the density of the hair cells. It is shown that these steps, when combined with skeleton correlation as a post-processing step, considerably improve the results for complex auditory events.

In addition to these improvements to the binaural auditory model, a new cognitive model is utilized which attempts to replicate the processes that are carried out by the human auditory system to estimate both the number of active sources as well as their direction. The combined system can then be used, e.g., as the preliminary stage of a common source separation method.

This paper is structured as follows: In Section 2, the general concept of the binaural hearing model and the novel extensions are described. In Section 3, the new cognitive strategy to estimate the number of sources is introduced followed by the experimental setup and a discussion of the results. Conclusive remarks are presented in Section 4.

2 Binaural Hearing Models

Mathematically speaking, the coincidence-based model is based on a short-time cross-correlation of the neural signals that are produced by the hair cells in the cochleas. A disadvantage of this model is that in the localization, only the ITD is taken into account. Given that our brain mostly utilizes the ILD to locate a sound event for frequencies greater than 1500 Hz, this is a major drawback.

To overcome this issue, Lindemann extended the model by two major points [10, 11]. First, monaural detectors are included in the model to ensure that the model output is sensible for monaural or near-monaural cases (i.e., the level of the signal at one ear is negligible compared to the other one). The most important extension, however, is an inhibition mechanism that suppresses the fire rate caused by coincidence units if the fire rate of neighboring coincidence units is very high. This so-called *contralateral inhibition* leads to sharper peaks of the correlation diagram along the τ axis, as well as increasing sensitivity to ILD. Moreover, the ambiguities about the ITD at higher frequencies are suppressed. In the next section, we briefly review the extended model. A detailed description is beyond the scope of this paper but can be found in [10, 11].

2.1 Hearing Model According to Lindemann

In this more detailed description of the binaural hearing model and the proposed extensions, the transfer functions of the outer and middle ear are not considered. Hence, the modeling process begins at the inner ear which was described in the original proposal as a 36-channel filter bank of band pass filters according to the critical bands [12].

These band pass filters cover the frequency range from 20 Hz to over 16 kHz. As an improvement to the original model, the filter bank here is implemented as a Gammatone filter bank which is known to be better adapted to the transmission function of the cochlea [13].

A post-processing stage that consists of a half-wave rectification, a square root function and a low pass filter with a cutoff frequency of 800 Hz completes the transfer function of the hair cells within the cochlea. This allows to extract the envelope of the stimulus which is needed to correctly model the capability of the human hearing system to utilize the ITD for the localization of complex stimuli if they exhibit a low frequency envelope.

The output of this stage (i.e., the signals $r(-M,n)$ and $l(M,n)$ with $\pm M$ as the extreme values on the discretized τ-axis and n as a discrete time index) is the input to the Lindemann model which consists of a bi-directional chain of delay elements $\Delta\tau$ with additional time and amplitude variable multipliers $(\otimes)$. A block diagram of the model is depicted in Fig. 1. The leftmost and rightmost channels therein are the aforementioned monaural detectors which are activated by sound events at one ear for which no corresponding event at the other ear is present.

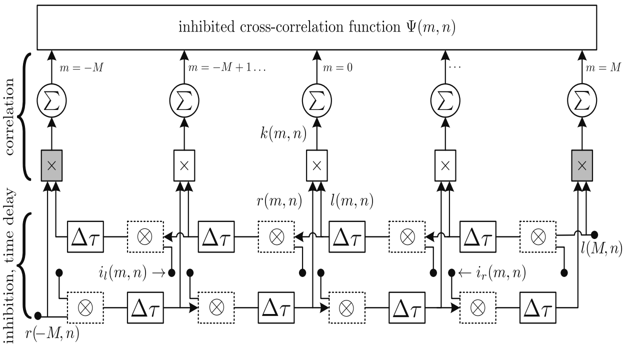

Figure 1: Block diagram of the binaural hearing model according to Lindemann [10]

The output of the model is the inhibited cross-correlation function $\Psi(m,n)$ which is calculated as

$$\Psi(m,n) = \sum_{\lambda=1}^{36} \sum_{i=n-N+1}^{n} k_\lambda(m,i)\, e^{\frac{i-n}{N}} \tag{1}$$

with the frequency band index λ and N as the number of samples per 5 ms frame. The cross-correlation products $k_\lambda(m,n)$ are obtained from the inhibited right and left signals $r_\lambda(m,n)$ and $l_\lambda(m,n)$ as follows:

$$k_\lambda(m,n) = r_\lambda(m,n) \cdot l_\lambda(m,n) \tag{2}$$

The inhibition mechanism according to [10, 11] has both a stationary and a dynamic component and is given by $i_r(m,n)$ and $i_l(m,n)$ in Fig. 1 for the right and left channel, respectively. The stationary inhibition decreases the amplitude of both signals before every delay element $\Delta\tau$ along the τ-axis with respect to the contralateral signal. The dynamic inhibition is a nonlinear lowpass whose input signal is the cross-correlation product $k(m,n-1)$. This dynamic inhibition is necessary in order to define further binaural properties as, e.g., the law of the first wave front.

With this inhibition mechanism, the model is also sensitive to ILD which leads to a much more realistic representation of the human capabilities. However, this can be disadvantageous if the signal amplitudes at the ears are significantly different. In that case, additional peaks are generated along the τ-axis [14]. A novel solution to overcome this drawback is shown in Section 2.2.

2.2 Improved delay and frequency weighting

In order to specifically suppress the additional peaks that stem from the monaural detectors, Gaik [14] already presented an additional weighting of the signal and extended the model by an adaptation to the individual head related transfer function (HRTF). After an extensive learning phase this leads to a more natural combination of ITD and ILD with the disadvantage that this only works well for the HRTFs that were used in the training phase.

A novel solution is proposed here that aims at correctly modeling the behaviour of the human auditory system with respect to the direction-dependent localization blur and the different importance of different frequencies for localization. The localization blur of the human auditory system is a function of the source direction of the sound in the horizontal plane: It ranges from just a few degrees in the front up to 10 degrees to the side [3]. It can be shown that the density of the coincidence counter units is similar to a Gaussian shape [15]. Hence, the inhibited cross-correlation function shall be weighted by the weighting factor $q_1(m)$ according to:

$$q_1(m) = \frac{5}{3\cdot\sqrt{2\pi}} \cdot e^{-\frac{25}{18}\cdot\left(\frac{m\cdot\Delta\tau}{\mathrm{ms}}\right)^2} \tag{3}$$

It has to be mentioned that both this internal time weighting and Gaik's proposal lead to more centrality, i.e., in situations where sources of similar intensity are active at the same time, the models will favor the centermost source. However, this is in agreement with the localization blur of the auditory system.

A further improvement can be achieved using a frequency weighting taking into account which frequencies have more significant contributions to the localization accuracy. Raatgever [16] has shown that especially the frequencies around the dominant region of 600 Hz are more important for localizing. From this fact, a weighting function $q_2(f)$ can be derived to weight the contributions from different subbands [17]:

$$q_2(f) = \begin{cases} 10^{-\left(\Sigma_{i=1}^{3} b_i\cdot f^i\right)/10} & f < 1200\,\mathrm{Hz} \\ 10^{-\left(\Sigma_{i=1}^{3} b_i\cdot 1200^i\right)/10} & f \geq 1200\,\mathrm{Hz} \end{cases} \tag{4}$$

with f in Hz and

b_1	b_2	b_3
$-9.383\cdot10^{-2}$	$1.126\cdot10^{-4}$	$-3.992\cdot10^{-8}$

These enhancements can be jointly integrated into Eq. 1 by a weighting function $q(m,f) = q_1(m)\cdot q_2(f)$ with the center frequency $f_m(\lambda)$ of the frequency band λ as follows:

$$\Psi(m,n) = \sum_{\lambda=1}^{36} \sum_{i=n-N+1}^{n} k_\lambda(m,i)\, e^{\frac{i-n}{N}} \cdot q(m, f_m(\lambda)) \tag{5}$$

2.3 Evaluation

The use of the aforementioned enhancements leads to more accurate neural activity patterns (NAPs). The impact of the improvements to the binaural model is illustrated by means of the resulting binaural excitation patterns for a complex sound event, both without (Fig. 2) and with (Fig. 3) the proposed model extensions. The example consists of a real world recording of two male English speakers, one stationary at an angle of 30 ° on the left and one moving from 60 ° on the right to 0 °.

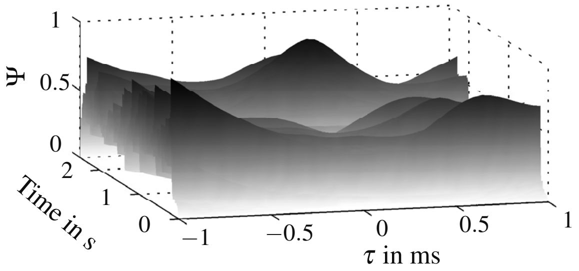

Figure 2: Binaural excitation pattern (original model)

When using the original model, a lot of the activities can be seen in the area of the monaural detectors (i.e., at values for τ of ±1 ms). Note that even though there is no simple relation between the τ-axis and the source direction, values for τ between ±0.6 ms approximately represent natural source directions while smaller and larger internal times indicate monaural signals.

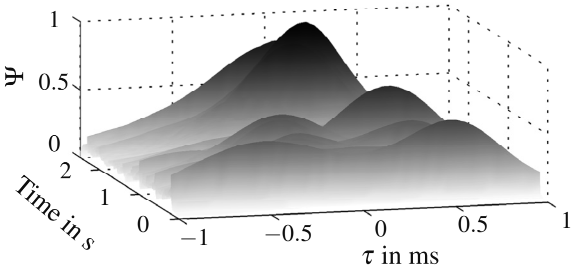

Figure 3: Binaural excitation pattern (improved model)

With the proposed improvements in place, the shape of the binaural excitation pattern is altered clearly and natural source directions are emphasized. As an additional pre-processing step for the blind clustering which will be described in more detail in Sec. 3, the resulting binaural excitation pattern is subject to a maximum search in every time interval leading to the skeleton cross-correlogram. The output of this step in this example is depicted in Fig. 4.

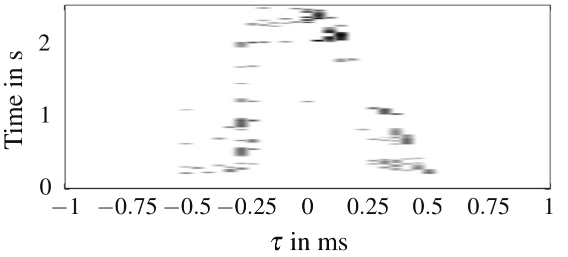

Figure 4: Skeleton cross-correlogram (improved model)

3 Blind Clustering

Looking at Fig. 4, it is obvious to the human eye that there is one stationary source at $\tau \approx -0.25$ms as can be seen by the vertical line of maxima there while a moving source is visible as the diagonal line from $\tau \approx 0.5$ms to $\tau \approx 0$ms in the diagram. Based on the improved hearing model, a novel cognitive processing scheme is proposed which can be used to estimate both the number of active sources as well as their direction.

3.1 Concept and Algorithm

The human brain is capable of separating different sources by identifying groups within the NAPs. The cognitive processing works in a very similar manner by applying *k-means clustering* [18] to the skeleton cross-correlogram using a Euclidean distance measure.

The distribution of the peaks in the skeleton cross-correlogram is depicted in the upper right part of Fig. 5. Each circle represents one maximum from Fig. 4 that is above a threshold of 20% of the amplitude of the highest maximum. Below this plot, a histogram of the distributions of τ is shown that is used to initialize the clustering algorithm: The number of local maxima in the histogram is chosen as the number of clusters while the positions of the maxima are chosen as the initial centroids for the clusters.

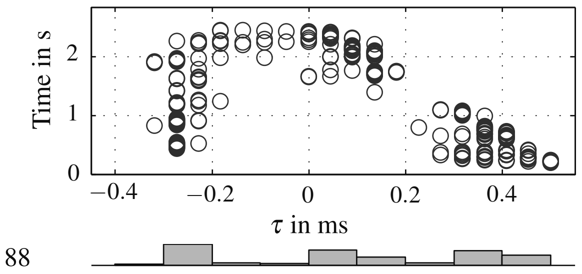

Figure 5: Skeleton cross-correlogram and derived histogram

In reverberant environments or for moving sources, this initialization can overestimate the number of active sources due to the fact that usually multiple local maxima appear in the histogram. In the presented example, three local maxima are found leading to the assumption that three sources are active. However, the maxima in the range of 0.05 and 0.35 ms belong to one moving source.

In order to accurately resolve this problem, a new refinement step is presented in Fig. 6. Based on the initial clustering (e.g., the three clusters ①, ② and ③ in this example as illustrated in Fig. 7), temporal and spectral information is extracted. The temporal information consists of indicators about the presence of simultaneous or only successive neural activity in different clusters. This information is derived from a comparison of the temporal spread of the initial clusters in the skeleton cross-correlogram. In order to compare the spectral properties of individual clusters, a time domain signal is constructed for each initial cluster by combining all audio signal segments that are associated with the NAPs of this cluster. After that, a spectral analysis of these time domain signals is performed by a single-level discrete wavelet transformation (DWT) with a Daubechies wavelet. Based on the output of the DWT, the clusters are examined for similarities by calculating the cross-correlation coefficients between the clusters.

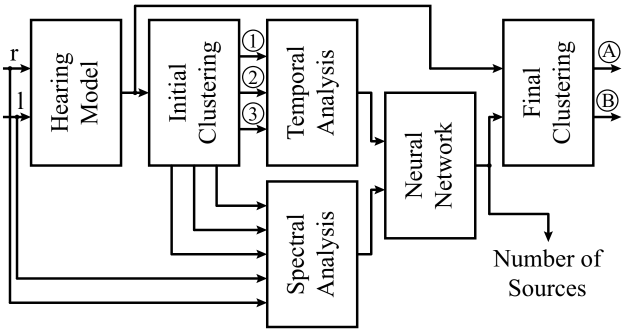

Figure 6: Refinement step for the estimation of the number of active sources

The temporal and spectral information is fed into a two-layer feedforward artificial neural network to determine the correct number of active sources. This network was trained by a supervised learning procedure (backpropagation) with a dataset of over 40 different complex sounds. The corrected number of clusters is fed back into the clustering process which can then exploit this additional information in order to improve the clustering performance.

3.2 Experimental Results and Limitations

The output of the initial clustering can be seen in the upper plot in Fig. 7. The stationary source is correctly identified as cluster ① but the moving source is separated into two clusters ② and ③. With the refinement step, these two clusters are joined to one new cluster Ⓑ as depicted in the lower plot in Fig. 7 while the stationary source, also a male English speaker, is still correctly identified as another cluster Ⓐ.

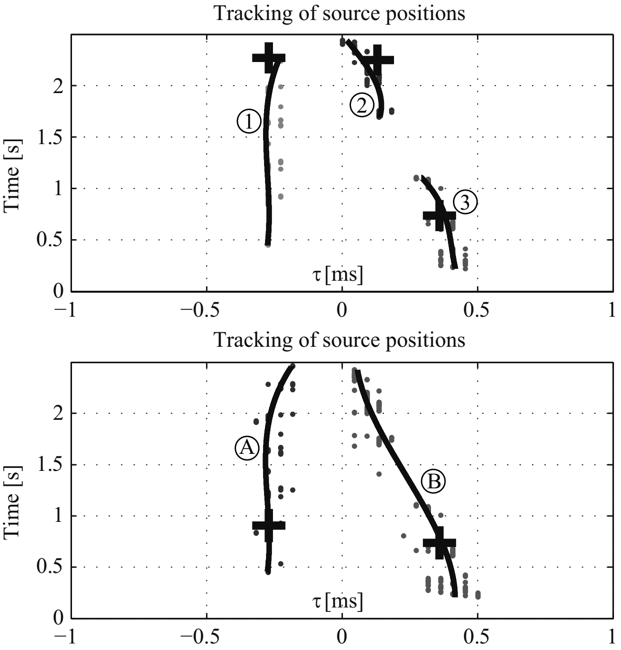

Figure 7: Initial (top) and corrected (bottom) clustering

While the model performs very well under various conditions, some limitations have to be mentioned. The robustness of the clustering decreases in very reverberant environments and for sources that are closer than 10 ° to each other. Additionally, since the model has neither the help of visual information nor the movement of the head, it is not able to resolve the ambiguity on the cone of confusion [3].

4 Conclusions

An advanced binaural auditory model and a mathematical modeling of cognitive behavior was proposed and shown to be able of estimating both the position and the number of acoustic sources. The model works reliably even for multiple concurrent speakers and moving targets. Possible application scenarios for the modeling scheme include all signal processing schemes that rely on an accurate representation of the human hearing system as well as source separation algorithms that require knowledge about the number of active sources.

References

[1] B. R. John William Strutt *Philosophical Magazine*, vol. 13, pp. 214–232, 1907.

[2] H. Fastl and E. Zwicker, *Psychoacoustics: facts and models*. Springer series in information sciences, Springer, 2007.

[3] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. The MIT Press, 1997.

[4] M. Bodden, “Binaural modeling and auditory scene analysis,” *Proceedings of WASPAA*, pp. 31–34, 1995.

[5] L. A. Jeffress, “A place theory of sound localization,” *J. Comp. Physiol. Psych.*, vol. 41, pp. 35–39, 1948.

[6] N. I. Durlach, “Equalization and cancellation theory of binaural masking-level differences,” *J. Acoust. Soc. Am.*, vol. 35, no. 8, pp. 1206–1218, 1963.

[7] T. C. Yin and J. C. Chan, “Interaural time sensitivity in medial superior olive of cat,” *J. Neurophysiol.*, vol. 64, pp. 465–488, 1990.

[8] H. Colburn and N. I. Durlach, *Hearing*, ch. 11, pp. 467–518. Academic Press, New York, 1978.

[9] J. Blauert, *Communication Acoustics*. Springer, 2005.

[10] W. Lindemann, “Extension of a binaural cross-correlation model by contralateral inhibition. I. Simulation of lateralization for stationary signals,” *J. Acoust. Soc. Am.*, vol. 80, no. 6, pp. 1608–1622, 1986.

[11] W. Lindemann, “Extension of a binaural cross-correlation model by contralateral inhibition. II. The law of the first wave front,” *J. Acoust. Soc. Am.*, vol. 80, no. 6, pp. 1623–1630, 1986.

[12] E. Zwicker and R. Feldtkeller, *Das Ohr als Nachrichtenempfänger*. S. Hirzel Verlag, 1967.

[13] J. Holdsworth and I. Nimmo-Smith, “Implementing a GammaTone Filter Bank,” tech. rep., Cambridge Electronic Design, MRC Applied Psychology Unit, 1988.

[14] W. Gaik, “Combined Evaluation of Interaural Time and Intensity Differences: Psychoacoustical Results and Computer Modeling,” *J. Acoust. Soc. Am.*, vol. 94, pp. 98–110, 1993.

[15] T. M. Shackleton, R. Meddis, and M. J. Hewitt, “Across frequency integration in a model of lateralization,” *J. Acoust. Soc. Am.*, vol. 91, no. 4, pp. 2276–2279, 1992.

[16] J. Raatgever, *On the Binaural Processing of Stimuli with Different Interaural Phase Relations*. PhD thesis, Technische Hogeschool Delft, The Netherlands, 1980.

[17] R. M. Stern and A. S. Zeiberg, “Lateralization of complex binaural stimuli: A weighted-image model,” *J. Acoust. Soc. Am.*, vol. 84, no. 1, pp. 156–165, 1988.

[18] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.