Azimuthal Localization of Concurrent Speakers Employing Interaural Coherence

M. Dietz, S. D. Ewert, V. Hohmann

Medizinische Physik, Universität Oldenburg, 26111 Oldenburg, Germany, Email: mathias.dietz@uni-oldenburg.de

Introduction

The human auditory system has a remarkable localization performance even in acoustically adverse conditions such as cocktail parties. The most robust localization cues are the interaural time difference (ITD) and the interaural level difference (ILD) both resulting from the azimuthal position of the sound source.

Technical algorithms for estimating the azimuthal direction of arrival (DOA) are also based on these parameters [1]. Most algorithms focus on ITDs for two main reasons. First ILDs can only be exploited with an obstacle between the sound receivers and second ITDs tend to be more robust against disturbances. In contrast to the human auditory system the common generalized cross-correlation algorithm [1] derives the ITD via Fourier transform of the complete frequency range.

In this contribution a localization model is presented which is motivated by the human auditory system. Interaural disparities are derived independently in several auditory filters resulting from an established model of auditory pre-processing [2].

The next section introduces the model before several results are presented in the subsequent section.

The auditory localization model

The model (Fig. 1) is based on the IPD model [3] which is a physiological plausible and psychoacoustically tested binaural model. In contrast to most binaural models which extract the ITD directly [e.g. 4, 5] it extracts the interaural phase difference (IPD) independently from the temporal fine structure (IPD^{fine}) and the stimulus envelope (IPD^{mod}). Nevertheless most of the results do not critically depend on the specific model of the binaural process and can also be achieved with conventional binaural models.

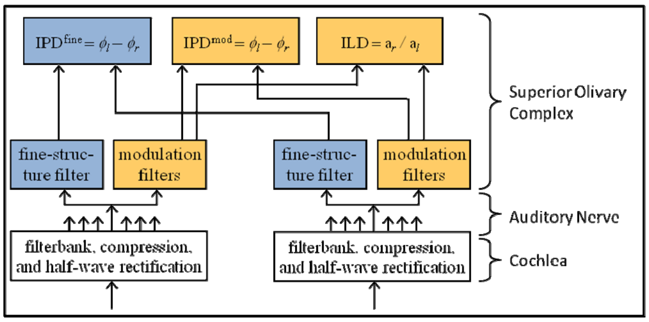

Figure 1: Processing scheme of the IPD model. After a gammatone filterbank a half-wave rectification spectrally separates the temporal fine-structure from the stimulus envelope. Both spectral regions are filtered and processed in separate pathways which can be assigned to the medial superior olive (IPD^{fine}) and the lateral superior olive (ILD and IPD^{mod}).

While IPD^{fine} is the most robust cue for localization it is ambiguous at frequencies above about 700 Hz for humans. Above this frequency the cycle duration is shorter than the ITD range observed for lateral sound sources between both ears and the IPD corresponds to several azimuthal angles. The model uses the ILD in order to resolve this ambiguity. The sign of the phase is set to the sign of the ILD. By this unwrapping the unambiguous region can be increased from [-π ; +π] to [-2π ; +2π] and the relevant frequency range is extended from 700 to about 1400 Hz.

Beside the remaining ambiguity above 1400 Hz, humans cannot exploit these cues for higher frequencies anyway, due to the loss of phase-locking to the temporal fine-structure.

The frequency range for the following fine-structure based localization is 200-1400 Hz, resulting in 12 auditory bands with a spacing of the respective equivalent rectangular bandwidth. A peripheral compression is assumed with a power of 0.4. The ILD is extracted over a 2nd order low-pass filter with 30 Hz cut-off. All other filters are complex-valued gammatone filters (see [3] for reference).

For higher frequencies temporal disparities in the envelope become more prominent and can be extracted via IPD^{mod}.

In addition to the IPDs the interaural coherence (IC) is determined for each band by the vector strength

$$\text{IC}(t) = \frac{1}{\tau_s} \cdot \left| \int_0^\infty \mathrm{d}\tau e^{i \cdot \text{IPD}(t-\tau)} e^{-\tau/\tau_s} \right| ,$$

where τ_s denotes the constant for temporal integration. In this study τ_s depends on the centre frequency f_c of the respective gammatone filter and is set to

$$\tau_s = \frac{5}{f_c} ,$$

i.e., the integration time is five times the cycle duration of the respective center frequency. For the localization analysis only those IPDs are extracted which have an IC > 0.98. These reliable IPD points are dominated by a single source and usually result in a more robust DOA estimation [4]. IC from vector strength is almost identical to the usual IC definition over the running cross-correlation function.

IPDs are mapped to an azimuthal axis where -90° is assigned to the left, 0° to straight ahead and +90° to the right. All reliable IPD points from 12 frequency bands are collected in the azimuthal plane. A heuristic model groups the points to single speakers under the assumption that the next point of the same speaker is less than 20° away from the last. If two moving speakers get closer than 20° the model groups the points on the basis of the fundamental frequency, according to the idea of the position-pitch-plane [6]. If the pitch is also very similar permutation errors may occur.

Results

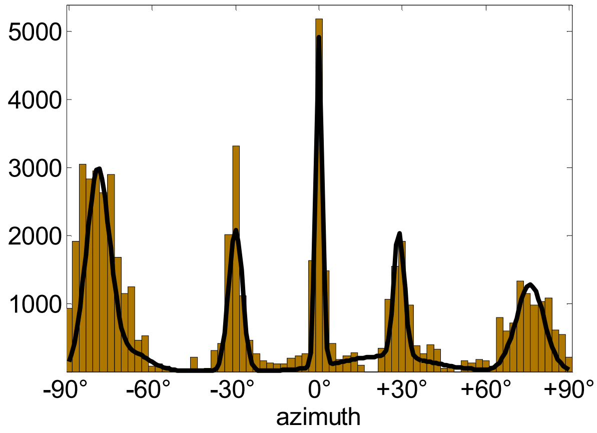

Figure 2: Histogram of reliable azimuth estimations for five simultaneous speakers at -80°, -30°, 0°, +30° and +80°. A Gaussian mixture model with 7 Gaussians finds the 5 dominant peaks at -80°, -30°, 0°, +29° and +76°.

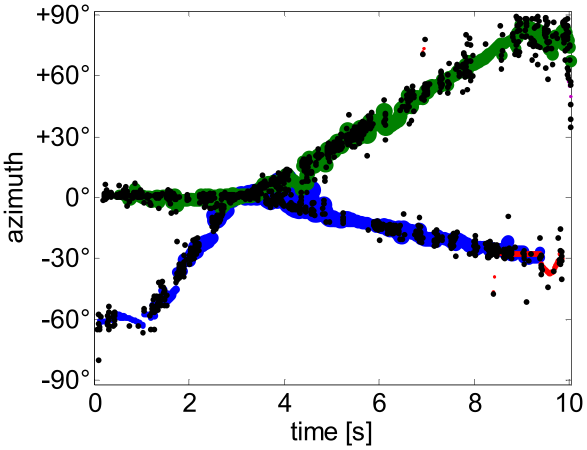

Figure 3: Tracking of two by-passing speakers. The dots indicate reliable azimuth estimations (high interaural coherence). The close by-passing is distinguished from a crossing due to differences in the fundamental frequencies. A third speaker is erroneously assumed after 8 sec at -40°. Line thickness is given by a leaky integration of the number of dots assigned to each speaker.

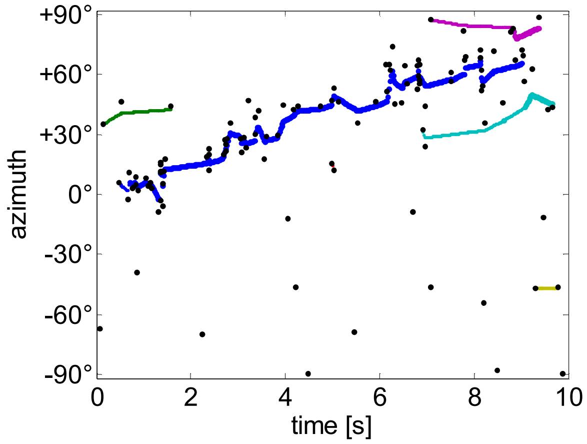

Figure 4: Tracking of one moving speaker (centre to right) in omnidirectional speech noise with a signal-to-noise ratio of 0 dB. Beside some misleading points the most reliable estimation coincides with the true speaker direction.

Discussion

The auditory localization model has successfully proven to cope with adverse conditions such as low signal-to-noise ratios or several concurrent speakers. A threshold interaural coherence for the DOA estimation was essential for this performance. In contrast to [4] the interaural coherence was derived with an f_c dependent temporal integration. This allows for a universal threshold for all frequency bands. Different time constants were tested and the best performance was achieved for short τ_s (10 ms at 500 Hz) comparable to [4] but much shorter than what is usually assumed for the "sluggish" binaural system (e.g. [5]).

As shown in [4] concurrent speakers are not found as secondary maxima in the cross-correlation function but rather as a temporally alternating position of a single main peak. Therefore it is not necessary to employ a complete axis of binaural coincidence detectors (delay-line). The sparse coding of the interaural phase difference by only one pair of neurons per band is sufficient to localize several speakers.

The independent processing of interaural disparities in each auditory filter offers advantages over integrated broad band processing. Spectrally differing sources can be separated easily and localized in different bands with little interaction. However, optimal combination of filters would be desirable, but is very task dependent.

Acknowledgments

This study was supported by the DFG (SFB/TRR 31 'The active auditory system'). We would like to thank the members of the Medical Physics group and Birger Kollmeier for continuous support and fruitful discussions.

References

[1] Kuhn, G. F. (**1977**). "Model for the interaural time differences in the azimuthal plane," J. Acoust. Soc. Am. **82**, 157-167.

[2] Dau, T., Püschel, D. and Kohlrausch, A. (**1996**). "A quantitative model of the 'effective' signal processing in the auditory system: I. Model structure," J. Acoust. Soc. Am. **99**, 3615-3622.

[3] Dietz, M., Ewert, S. D., Hohmann, V. and Kollmeier, B. (**2008**). "Coding of temporally fluctuating interaural timing disparities in a binaural processing model based on phase differences," Brain Res. **1220**, 234-245.

[4] Faller, C. and Merimaa, J. (**2004**). "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," J. Acoust. Soc. Am. **116**, 3075-3089.

[5] Breebaart, J., van de Par, S. and Kohlrausch, A. (**2001)**. "Binaural processing model based on contralateral inhibition. I. Model structure," J. Acoust. Soc. Am. **110**, 1074-1088.

[6] Képesi, M., Pernkopf, F. and Wohlmayr, M. (**2007**). "Joint position-pitch tracking for 2-channel audio," IEEE Trans. Content-Based Multim. Indexing. 303-306.