

Binaural Localization and Detection of Speakers in Complex Acoustic Scenes

T. May, S. van de Par and A. Kohlrausch

1 Introduction

The robust localization of speakers is a very important building block that is required for many applications, such as hearing aids, hands-free telephony, voice-controlled devices and teleconferencing systems. Despite decades of research, the task of robustly determining the position of multiple active speakers in adverse acoustic scenarios has remained a major problem for machines. One of the most decisive factors that influence the localization performance of algorithms is the number of microphones. When several pairs of microphones are available, beamforming techniques such as the *steered-response power*, SRP, approach [25] or the *multi-channel cross-correlation coefficient*, MCCC, method [7] can be applied to disambiguate the localization information by exploiting correlation among multiple pairs of microphones. Furthermore, high-resolution subspace techniques such as the *multiple signal classification*, MUSIC, algorithm [66] and the *estimation of signal parameters via rotational-invariance techniques*, ESPRIT, approach [64] generally require that the number of sensors is greater than the number of sound sources. *Blind source separation* approaches, such as the *degenerate unmixing estimation technique*, DUET, attempt to blindly localize and recover the signals of N sound sources from M microphone signals [38, 81]. Although the DUET system is able to deal with

T. May

Centre for Applied Hearing Research, Department of Electrical Engineering,
Technical University of Denmark, Kgs. Lyngby, Denmark

S. van de Par

University of Oldenburg, Oldenburg, Germany

A. Kohlrausch (✉)

Eindhoven University of Technology, Eindhoven, The Netherlands
e-mail: a.kohlrausch@tue.nl

A. Kohlrausch

Philips Research Europe, Eindhoven, The Netherlands

underdetermined mixtures, that is, $N > M$, in anechoic conditions, performance deteriorates in reverberant environments.

In contrast to machines, the human auditory system is remarkably robust in complex multi-source scenarios. It can localize and recognize up to six competing talkers [12], in spite of the fact that it is provided with only two signals reaching the left and the right ears. Moreover, listening with two ears substantially contributes to the ability to understand speech in multi-source scenarios [11, 19]. Unlike blind source separation algorithms that aim at separating the sources in such a way that they are fully reconstructed, the human auditory system does not need to perform such a reconstruction of the original signals. It only needs to extract those properties of the signal of interest that are needed for a particular task, such as estimating the direction of a sound source, the identity of a speaker, or the words that are being pronounced. Thus, when particular parts of the target signal are not available, that is, *missing*, due to the presence of other interfering sources, there may still be enough information, in other words, perceptual cues, available to extract the properties of interest, for example, the identity of a speaker. This ability of the human auditory system to handle complex multi-source scenarios and to segregate the contributions of individual sound sources is commonly summarized by the term *auditory scene analysis*, ASA. As described by Bregman [10], the underlying principles that facilitate ASA can be divided into two stages, namely, segmentation and grouping. First, the acoustic input is decomposed into spectro-temporal units, where each individual unit is assumed to be dominated by one particular source. Secondly, in the grouping stage, a set of primitive grouping rules, termed *Gestalt principles*, are employed by the auditory system in order to integrate the information that is associated with a single sound source. These Gestalt principles can be considered as data-driven mechanisms that are related to physical properties of sound generation, leading to certain structures in auditory signals. Common onsets across frequency, common amplitude and frequency modulation, and common spatial location are examples of such Gestalt principles [10, 23, 75]. Apart from data-driven processing—also known as *bottom-up processing*—the auditory system is able to focus the attention on a particular target source and interpret the underlying source, for instance, in order to understand speech. This involves schema-driven processing—also referred to as *top-down processing*—and requires *a priori* knowledge about different sound sources.

Inspired by the robustness of the human auditory system, a research field termed *computational auditory scene analysis*, CASA, has emerged, which aims at reproducing the capabilities of the human auditory system with machines on the basis of sensory input [75]. As the analysis is restricted to binaural signals, the task of automatically localizing multiple competing sound sources is particularly challenging. In this chapter, only two microphone signals will be considered, corresponding to the left- and the right ear signals of an artificial head, and it is shown how principles of human auditory processing can be used to estimate the azimuth of multiple speakers in the presence of reverberation and interfering noise sources, where the number of active speakers is assumed to be known *a priori*. Note that the intention is to develop a robust computer algorithm that is inspired by auditory mechanisms, rather than building a physiologically-plausible model of the human auditory system. Although

this chapter focuses on binaural signals, the presented approach can be extended to microphone arrays with multiple pairs of microphones.

After describing the binaural signals that are used throughout this chapter, an overview of different approaches to binaural sound-source localization, ranging from technical approaches to auditory-inspired systems, will be given in Sect. 3. A thorough analysis of localization performance will then be presented in Sects. 4 and 5, using multiple competing speakers in reverberant environments. An important problem is the influence of noise on speaker-localization performance, which will be discussed in Sect. 6. In particular, it will be shown that the ability to localize speakers is strongly influenced by the spatial diffuseness of the interfering noise. Moreover, it will be seen that the presence of a compact noise source imposes severe challenges for correlation-based approaches. By employing principles of auditory grouping based on common spatial-location and missing data classification techniques, it is possible to make a distinction between source activity that originates from speech- or from noise sources. This distinction can substantially improve the speaker-localization performance in the presence of interfering noise.

2 Simulation of Complex Acoustic Scenes

In order to evaluate the localization algorithms that are presented in this chapter, complex acoustic scenes are simulated by mixing various speech and noise sources that are placed at different positions within a room. Binaural signals are obtained by convolving monaural speech files with binaural room impulse responses, BRIRs, corresponding to a particular sound-source direction. These BRIRs are simulated by combining a set of head-related transfer functions, HRTFs, with room impulse responses, RIRs, that are artificially created according to the image-source model [4]. More specifically, the MIT database is used, which contains HRTFs of a KEMAR¹ artificial head that were measured at a distance of 1.4 m in an anechoic chamber [30]. These HRTFs are combined with RIRs, simulated with *ROOMSIM*,² a MATLAB toolbox provided by Schimmel et al. [65]. The receiver, KEMAR, was placed at seven different positions in a simulated room of dimensions $6.6 \times 8.6 \times 3$ m. For the experiments conducted in this chapter, a set of BRIRs with the following reverberation times are simulated for each of the seven receiver positions, namely, $T_{60} = \{0.2, 0.36, 0.5, 0.62, 0.81 \text{ and } 1.05\}$ s. The reverberation time, T_{60} , of the simulated BRIRs has been verified by applying the energy-decay-curve method developed by Schroeder [67].

Furthermore, a number of databases with measured BRIRs are publicly available [35, 37, 39], each of them focusing on a particular application. For a systematic

¹ Knowles electronic manikin for acoustic research, KEMAR.

² Although the problem of moving sources is not covered in this chapter, the MATLAB toolbox *ROOMSIMOVE* for simulating RIRs for moving sources can be found at <http://www.irisa.fr/metiss/members/evincent/software>.

analysis of localization performance, the measurements provided by the University of Surrey [35] were selected, since they offer BRIRs recorded in four different rooms with an azimuthal resolution of 5° . The following set of measured BRIRs is used for evaluation, $T_{60} = \{0.32, 0.47, 0.68 \text{ and } 0.89 \text{ s}\}$. Note that the BRIRs of the Surrey database are recorded with a Cortex–MK.2 head-and-torso simulator, HATS, which is different from the KEMAR artificial head that was used to create the simulated BRIRs. This allows the investigation of the impact on localization performance that is induced by a mismatch between BRIRs that are used for training and those which are used for testing. The results will be reported in Sect. 5.

Multi-source mixtures are created by randomly positioning sound sources within the azimuth range of $[-90, 90^\circ]$ while having an angular distance of at least 10° between neighboring sources. For the experiments presented in Sects. 4 and 5, speech files are randomly selected from the speech-separation challenge, SSC, database [22]. Signals are either trimmed or concatenated to match an overall duration of 2 s. The level of multiple competing speech sources was always set equal. In addition, the impact of interfering noise on localization performance is systematically investigated in Sect. 6 by using three different types of noise signals, namely, babble noise and factory noise from the NOISEX database [74] and speech-shaped noise that is based on the long-term average spectrum, LTAS, of 300 randomly-selected speech files. Interfering noise sources are simulated by randomly selecting different time segments of the corresponding type of background noise. In contrast to the speech files, there is no constraint on the angular distance between multiple noise sources. The signal-to-noise ratio, SNR, is adjusted by comparing the energy of all speech sources to the energy of the noise. Note that the energies of the left and the right signals are added prior to SNR calculation. The resulting binaural multi-source signals are sampled at a sampling frequency of $f_s = 16 \text{ kHz}$.

3 Binaural Sound-Source Localization

The two major physical cues that enable human sound-source localization in the horizontal plane are *interaural time differences*, ITDs, and *interaural level differences*, ILDs, between the two ears [60]. Both cues are complementary in their effectiveness. As already formulated by Lord Rayleigh more than 100 years ago, the ITD cue is most reliable at low frequencies, whereas the ILD cue is more salient at higher frequencies [60]. The spectral modifications provided by the complex shape of the external ears are particularly important for the perception of elevation and help to resolve front-back confusions [68]. In this chapter, the localization of sound sources is restricted to the frontal horizontal plane within the area of $[-90, 90^\circ]$. In the following sections, a short review of popular sound-source localization approaches will be given with the special application to binaural signals.

3.1 Broadband Approaches

One of the most frequently-used approaches to sound-source localization is to estimate the time difference of arrival, TDOA, between a pair of two spatially separated microphones. This approach usually consists of the following two steps. First, the relative delay between the microphones is estimated. Secondly, the estimated delay is used to infer the actual angle of the sound source by employing knowledge about the microphone-array geometry.

The *generalized cross-correlation*, GCC, framework presented by Knapp and Carter [41] is the most popular approach to perform time-delay estimation. The TDOA estimate, $\hat{\tau}$, in samples, is obtained as the time lag, τ , that maximizes the cross-correlation function between the two filtered microphone signals, that is,

$$\hat{\tau} = \arg \max_{\tau} \frac{1}{2\pi} \int_{\omega} W(\omega) X_L(\omega) X_R^*(\omega) e^{j2\pi\omega\tau} d\omega, \quad (1)$$

where $X_L(\omega)$ and $X_R(\omega)$ indicate the short-time Fourier transforms of the microphone signals, $x_L(n)$ and $x_R(n)$, received at the left and the right ears, and $W(\omega)$ denotes a frequency-dependent weighting function. The classical cross-correlation, CC, method uniformly weights all frequency components by setting $W_{CC}(\omega) = 1$. To increase the resolution of GCC-based time delay estimation, it is useful to interpolate the GCC function by an oversampled inverse fast Fourier transform, IFFT [27]. Hence, an oversampling factor of four is considered in this chapter, resulting in a τ -step size of 16 μ s.

In ideal acoustic conditions, in which the signals captured by the two microphones are simply time-shifted versions of each other, the most prominent peak in the GCC function reveals the true TDOA between both microphones and can be reliably detected. However, in more realistic scenarios with reverberation and environmental noise, the identification of peaks in the GCC function becomes less accurate, which, in turn, reduces the localization performance. Therefore, a variety of different weighting functions have been proposed in order to sharpen the peak that corresponds to the true TDOA and to improve its detectability [15, 41]. Among them, the so-called *phase transform*, PHAT, is the most frequently-used weighting function, which whitens the cross-spectrum between the two microphone signals, $x_L(n)$ and $x_R(n)$, prior to cross-correlation by choosing the weighting as $W_{PHAT}(\omega) = |X_L(\omega) X_R^*(\omega)|^{-1}$. When ignoring the impact of noise, the PHAT weighting eliminates the influence of the source signal on localization and exhibits a clearly visible peak at the true TDOA. One apparent drawback of the PHAT weighting is that it gives equal weight to all frequencies, regardless of their signal-to-noise ratio, SNR. Nevertheless, if all interferences can be attributed to reverberation, the PHAT weighting has been shown to achieve robust localization performance [32, 83], as long as the level of noise is low [83].

Another approach that attempts to improve the robustness of the GCC function in noisy and reverberant environments is to perform *linear prediction*, LP, analysis to

extract the excitation source information³ [59]. The conventional GCC function is then computed based on the Hilbert envelope of the LP-residual signal, which was reported to form a more prominent main peak at the true TDOA in comparison to the conventional CC weighting.

Alternatively, the delay can also be derived from the *average magnitude-difference function*, AMDF, and its variations [17, 36]. For an comprehensive overview of different time-delay-estimation techniques the reader is referred to Chen et al. [18].

Once an estimation of the time delay between the left and the right ears is available, the second step of TDOA estimation requires conversion of the measured time delay to its corresponding *direction of arrival*, DOA. This is commonly achieved by a table-look-up procedure that can roughly account for the diffraction effects of the human head. Therefore, the estimated delay, $\hat{\tau}$, of a particular TDOA method is monitored in response to white noise filtered with HRTFs that are systematically varied between -90° and 90° [8, 57]. The resulting mapping function establishes a monotonic relation between time delay, $\hat{\tau}$, and sound-source azimuth, φ , at an angular resolution of 1° .

3.2 Auditory-Inspired Approaches

It is an important property of the human auditory system to be able to segregate the individual contributions of competing sound sources. In an attempt to incorporate aspects of peripheral auditory processing, the cross-correlation analysis can be performed separately for different frequency channels [8, 46, 50, 57, 63]. The frequency selectivity of the basilar membrane is commonly emulated by a *Gammatone filterbank*, GTFB, that decomposes the acoustic input into individual frequency channels with center frequencies equally spaced on the *equivalent-rectangular-bandwidth-rate* scale, ERB scale, [31]. It is advantageous to use phase-compensated Gammatone filters by accounting for the frequency-dependent group delay of the filters at their nominal center frequencies, c_f . This time-alignment can be achieved by introducing a channel-dependent time lead and a phase-correction term [14], allowing for a synchronized analysis at a common instance of time. Further processing stages crudely approximate the neural-transduction process in the inner hair cells by applying half-wave rectification and square-root compression to the output of each individual Gammatone filter [63]. Although not considered in this chapter, more elaborate models of the neural-transduction process might be applied at this stage [53, 54, 72]. Then, on the basis of these auditory signals, denoted as $h_{L,f}$ and $h_{R,f}$, the normalized cross-correlation, C , can be computed over a window of B samples as a function of time lag, τ , frame number, t , and frequency channel, f , as follows,

³ The corresponding MATLAB code can be found at http://www.umiacs.umd.edu/labs/cvl/pirl/vikas/Current_research/time_delay_estimation/time_delay_estimation.html

$$C(t, f, \tau) = \frac{\sum_{i=0}^{B-1} (h_{L,f}(t \cdot B/2 - i) - \bar{h}_{L,f}) (h_{R,f}(t \cdot B/2 - i - \tau) - \bar{h}_{R,f})}{\sqrt{\sum_{i=0}^{B-1} (h_{L,f}(t \cdot B/2 - i) - \bar{h}_{L,f})^2} \sqrt{\sum_{i=0}^{B-1} (h_{R,f}(t \cdot B/2 - i - \tau) - \bar{h}_{R,f})^2}} . \quad (2)$$

$\bar{h}_{L,f}$ and $\bar{h}_{R,f}$ denote the mean values of the left and right auditory signals estimated over frame t . The normalized cross-correlation function is evaluated for time lags within a range of $[-1, 1]$ ms], and the lag that corresponds to its maximum is used to reflect the interaural time difference, ITD,

$$\widehat{\text{itd}}(t, f) = \arg \max_{\tau} C(t, f, \tau) / f_s . \quad (3)$$

Instead of using the integer time lag directly for ITD estimation, it is possible to refine the fractional peak position by applying parabolic [36] or exponential [84] interpolation strategies. It has been found that the exponential interpolation performs better than the parabolic one, which is in line with results reported by Tervo and Lokki [73].

The frequency-selective processing allows the frequency-dependent diffraction effects introduced by the shape of the human head [8, 57, 63] to be accounted for. More specifically, the cross-correlation pattern, $C(t, f, \tau)$, which is usually a function of the time lag, τ , is warped onto an azimuth grid, $S(t, f, \varphi)$ [57]. This warping is accomplished by a frequency-dependent table look-up, which is obtained in a similar way as the one described in Sect. 3.1 and translates time delay to its corresponding azimuth. The frame-based source position can then be obtained by integrating the warped cross-correlation patterns across frequency and locating the most prominent peak in the summary cross-correlation function, that is,

$$\hat{\varphi}_{\text{GFB}}(t) = \arg \max_{\varphi} \sum_f S(t, f, \varphi) . \quad (4)$$

This across-frequency integration is an implementation of the *straightness approach* where sound-source directions with synchronous activity across multiple frequency channels are emphasized [69, 71].

If more than one sound source should be resolved on a frame-by-frame basis, it might be beneficial to compute a *skeleton* cross-correlation function [57, 63]. The general concept is that each local peak in the cross-correlation function is replaced by a Gaussian function where the corresponding standard deviation is varied linearly as a function of the frequency channel. This processing aims at sharpening the response of the summary cross-correlation function.

Although the computational approaches to binaural sound-source localization discussed so far have been focusing on exploiting the ITD cue, there are some attempts to also consider the information that is supplied by the interaural level differences, ILDs. The aforementioned skeleton cross-correlation function has some similarities with the concept of *contralateral inhibition*, where the ILD information is incorporated

into the cross-correlation framework to predict phenomena related to the *precedence effect* [44, 45]. A comprehensive review of the recent development of binaural models can be found in [9]. Moreover, the model presented by Palomäki et al. [57] uses an azimuth-specific ILD template to verify if the estimated ILD is consistent with the template ILD that is expected for the ITD-based azimuth estimate. The ILD cue can be derived by comparing the energy of the left- and the right-ear signals, $h_{L,f}$ and $h_{R,f}$, over a window of B samples, namely,

$$\widehat{\text{ild}}(t, f) = 10 \log_{10} \left(\frac{\sum_{i=0}^{B-1} h_{R,f}(t \cdot B/2 - i)^2}{\sum_{i=0}^{B-1} h_{L,f}(t \cdot B/2 - i)^2} \right). \quad (5)$$

3.3 Supervised-Learning Approaches

In many realistic environments the observed binaural cues will be affected by the presence of reverberation and noise sources. Although the binaural cues are noisy, there still is a certain degree of predictability associated with these binaural cues, depending on the azimuth of the sound source. Recently, supervised-learning strategies have been employed in order to optimally infer the location of a source on the basis of binaural cues [24, 34, 50, 56, 77–79] where the interdependence between interaural time and level differences can be jointly considered as a function of frequency channel, f , and sound-source direction, φ . Note that supervised-learning approaches based on binaural cues have also been applied in the context of sound-source segregation [33, 63].

In this chapter, a *Gaussian mixture model*, GMM, classifier to approximate the two-dimensional feature distribution of ITDs and ILDs will be described. For the extraction of ITDs and ILDs, auditory front-ends as described in the previous section are commonly employed. In contrast to utilizing a mapping function—see Sects. 3.1 and 3.2—that translates the obtained interaural differences to their corresponding sound-source directions, supervised-learning approaches offer the considerable advantage of providing a probabilistic framework where multiple layers of information can be jointly analyzed. This combined analysis of ITDs and ILDs has been shown to be superior to exclusively relying on the ITD cue [50]. The localization framework based on GMMs is very flexible and can be readily extended to incorporate additional features that depend on the sound-source direction. Likewise, the GMM framework is applicable to array geometries with more than two microphones from which binaural features for multiple microphone pairs could be extracted. To extend the working range of the GMM-based localization model to the dimension of elevation, the additional integration of monaural cues might be beneficial [43, 82]. Further details about vertical sound-source localization can be found in [6], this volume.

During the supervised training process, *a priori* knowledge is available to create training data, namely, binaural features, and the corresponding class labels that categorize the training data according to different sound-source directions, φ . As analyzed by Roman et al. [63], the joint distribution of ITDs and ILDs is influenced by the presence of a competing source and its strength relative to the target source. This can be accounted for by training the localization model with binaural cues extracted for mixtures with a target and an interfering source at various SNRs. The resulting model was reported to yield substantial SNR improvements [63], however, its application is restricted to anechoic scenarios.

Multi-Conditional Training of Binaural Cues

Localization models are commonly based on the assumption of single-path wave propagation. To overcome this fundamental limitation, a multi-conditional training stage can be applied in order to incorporate possible variations of ITDs and ILDs that are caused by the presence of competing sound sources, room reverberation and changes in the source-receiver configuration [50]. During the multi-conditional training stage, a variety of different acoustic conditions are simulated, and the frequency-dependent distributions of binaural features are approximated by a Gaussian mixture model classifier. The reverberation characteristic is intentionally simplified by assuming a frequency-independent reverberation time of $T_{60} = 0.5$ s. In this way, the same amount of uncertainty is encoded in each Gammatone channel. To ensure that the model is not trained for a particular room position, the multi-conditional training also involves various receiver positions and radial distances between the source and the receiver. Note that these positions are different from the ones that are used for evaluation—see Sect. 2 for details. More specifically, the following parameters are varied for each sound-source direction, φ ,

- Competing speaker at $\pm 40^\circ$, $\pm 30^\circ$, $\pm 20^\circ$, $\pm 10^\circ$ and $\pm 5^\circ$ relative to the azimuth φ of the target source
- Three SNRs between the target and the competing source, 20, 10 and 0 dB
- Three radial distances between the target source and the receiver, 0.5, 1 and 2 m
- Eight positions within the simulated room of dimensions, $6.6 \times 8.6 \times 3$ m

To visualize the influence of reverberation and the presence of multiple sound sources on ITDs and ILDs, the binaural feature space created by the multi-conditional training stage is presented in Fig. 1. Each dot represents a joint ITD-ILD estimate obtained for time frames of 20 ms. Note that the black and the gray distributions correspond to binaural cues associated with a target source at $\varphi = -50^\circ$ and $\varphi = 50^\circ$, respectively. When analyzing the general shape of the joint ITD-ILD feature distributions, it can be seen that the interdependency of both binaural cues results in complex multi-modal patterns. Due to spatial aliasing, the cross-correlation function leads to ambiguous ITD estimates at higher frequencies at which the wavelength is smaller than the diameter of the head. Consequently, the ambiguous ITD information results in multi-modal distributions where the number of individual clusters systematically

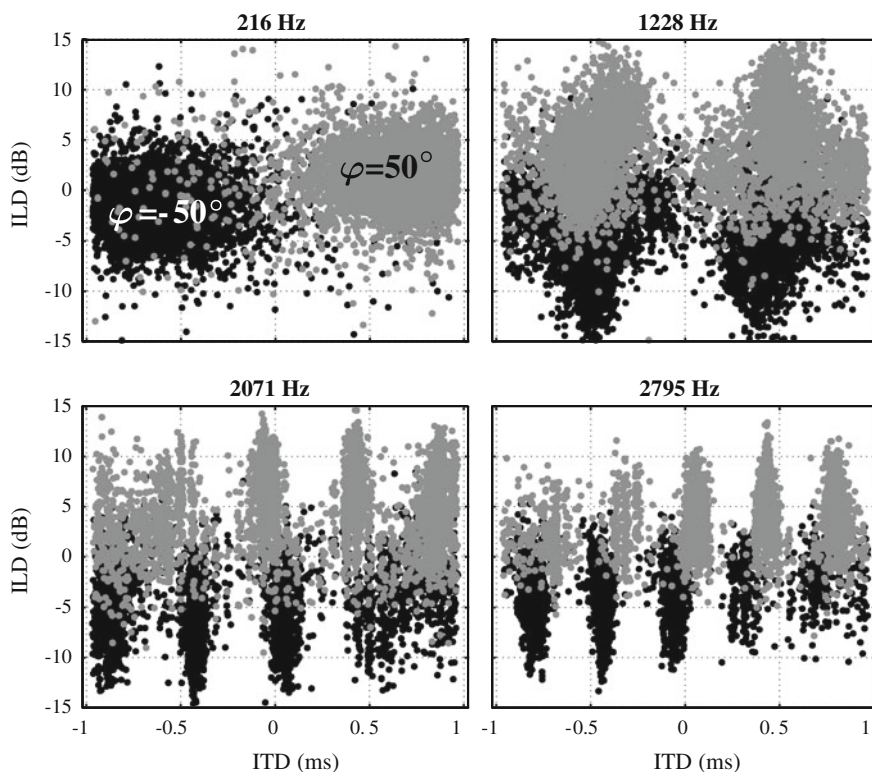


Fig. 1 Frequency-dependent distributions of interaural time and level differences, ITDs and ILDs, created by the multi-conditional training stage. Each dot represents a frame-based observation of the joint ITD and ILT feature space. The *black* and *gray* distributions correspond to two different sound-source directions, namely, $\varphi = -50^\circ$ and $\varphi = 50^\circ$ respectively. See text for more details

increases with frequency. This ITD fine structure at higher frequencies is deliberately maintained, because experiments showed that a more detailed hair-cell model that simulates the inability of the human auditory system to analyze the temporal fine structure at frequencies above 1.5 kHz performed substantially worse in terms of localization accuracy [50]. This comparison suggests that the fine-structure information of the ITD can be effectively exploited by the GMM classifier for improved localization performance. This is a distinguishing feature from other localization models that attempt to build a physiologically-plausible model of human sound-source localization [26]. Another practical advantage of exploiting ITDs at higher frequencies is that the reverberation energy usually decays towards higher frequencies. As a result, the binaural cues associated with higher frequencies are less affected by reverberation, and thus convey more reliable contributions to overall localization. The spread of the individual clusters can be attributed to the impact of reverberation and the presence of a competing source. Furthermore, when comparing the binau-

ral features for $\varphi = -50^\circ$ and $\varphi = 50^\circ$, it can be seen that the complex structure systematically shifts with sound-source direction.

GMM-Based Localization

The distinct change of ITDs and ILDs as a function of sound-source direction, which is illustrated in Fig. 1, can now be systematically learned by a GMM classifier. Thus, the multi-conditional training is performed for a set of $K = 37$ sound-source directions, $\{\varphi_1, \dots, \varphi_K\}$ spaced by 5° , within the range of $[-90, 90^\circ]$. After training, a set of frequency- and azimuth-dependent diagonal GMMs, $\{\lambda_{f,\varphi_1}, \dots, \lambda_{f,\varphi_K}\}$, is available. Given an observed binaural feature vector consisting of estimated ITDs and ILDs, $\mathbf{x}_{t,f} = \{\widehat{\text{itd}}(t, f), \widehat{\text{ild}}(t, f)\}$, the three-dimensional spatial log-likelihood can be computed for the k th sound-source direction being active at time frame t and frequency channel f as

$$\mathcal{L}(t, f, k) = \log p(\mathbf{x}_{t,f} | \lambda_{f,\varphi_k}) . \quad (6)$$

To obtain a robust estimation of sound-source direction, the log-likelihoods are accumulated across all frequency channels and the most probable direction reflects the estimated source location on a frame-by-frame basis, that is,

$$\hat{\varphi}_{\text{GMM}}(t) = \arg \max_{1 \leq k \leq K} \sum_{f=1}^F \mathcal{L}(t, f, k) . \quad (7)$$

Note that, in contrast to integrating the cross-correlation pattern across frequency, see (4), the log-likelihoods are accumulated, taking into account the uncertainty of binaural cues in individual frequency channels. This probabilistic integration of binaural cues has been also suggested by Nix and Hohmann [56]. As a result, the model does not require additional selection mechanisms, such as the coherence-based selection of reliable binaural cues [29], because this weighting is already implicitly incorporated into the model by the multi-conditional training stage. In other words, the multi-conditional training considers possible variations of interaural time and level differences resulting from competing sound sources and room reverberation, thus improving the robustness of the localization model in adverse acoustic scenarios.

4 Frame-Based Localization of a Single Source

In this section a comparison is performed of the ability of different approaches to localize the real position of one speaker in the presence of reverberation, based on 20 ms time frames. Therefore, binaural mixtures are created by using the simulated BRIRs with different reverberation times, T_{60} . A set of 185 binaural mixtures

is created for each reverberation time. For evaluation, the following methods are considered,

- Generalized cross-correlation, GCC, function according to (1) with two different weighting functions, W_{CC} and W_{PHAT}
- GCC function according to (1) with W_{CC} based on the LP residual
- Gammatone-based cross-correlation, GCC–GTFB, according to (4)
- GMM-based localization according to (7) with multi-conditional training

The GCC-based algorithms used a 20 ms Hamming window and a fast Fourier transform of 1,024 samples. The resolution of the resulting TDOA estimate was improved by applying an IFFT-based interpolation with an oversampling factor of four. The LP residual is created on the basis of 20 ms frames by using ten LP-filter coefficients. The Gammatone-based processing is based on 32 auditory filters that were equally distributed on the ERB-rate scale between 80 Hz and 5 kHz. All mapping functions are derived from anechoic BRIRs based on the KEMAR HRTFs. In general, the number of active target speakers is assumed to be known *a priori*. The blind estimation of the number of active speakers is currently being investigated [48].

The percentage of correctly localized frames is shown in Fig. 2 as a function of the absolute error threshold. Different panels represent different reverberation times, ranging from $T_{60} = 0.2$ up to $T_{60} = 1.05$ s. Apart from the conventional GCC approach, all algorithms reach ceiling performance for a moderate reverberation time of $T_{60} = 0.2$ s. But with increasing reverberation time, performance of all GCC-based methods substantially deteriorates. Due to the fact that these approaches are based on the assumption of single-path wave propagation, the presence of strong reflections causes spurious peaks in the GCC function that are erroneously selected as source positions. Thus, localization performance of the GCC-based approaches will inevitably decrease in more challenging acoustic conditions. While the LP-based preprocessing improves the performance of the conventional GCC approach, the PHAT-weighting produces the overall most reliable estimates of all GCC-based approaches, which supports the findings of previous studies [32, 83]. The GMM-based localization model shows superior performance, especially in conditions with strong reverberation, suggesting that the multi-conditional training stage can account for the distortions of ITDs and ILDs due to reverberation. Furthermore, unlike the other approaches, the GMM-based localization model is able to jointly analyze ITD and ILD information.

5 Localization of Multiple Sound Sources

In more complex acoustic scenarios, a variety of sound sources might be active at the same time. As demonstrated in the previous section, the performance of localizing only one speaker on a frame-by-frame basis noticeably degrades with increasing reverberation time. Therefore, an important question is how to integrate localization information across time in order to reliably resolve the position of multiple competing sound sources in reverberant environments.

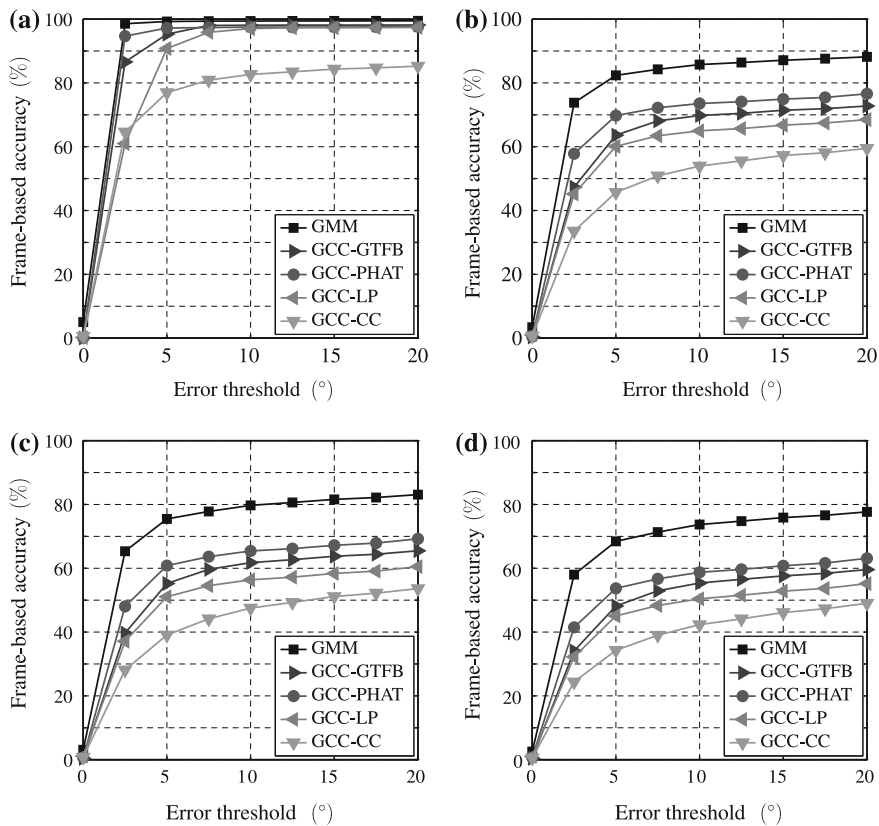


Fig. 2 Frame-based accuracy in % of localizing one speaker in a reverberant room as a function of the absolute error threshold in $^{\circ}$. Results are shown for different reverberation. **a** $T_{60} = 0.2$ s. **b** $T_{60} = 0.5$ s. **c** $T_{60} = 0.81$ s. **d** $T_{60} = 1.05$ s

Temporal Integration

Recursive smoothing techniques could be considered as a way to calculate a running average of localization information. Regarding the class of GCC-based approaches that require a short-time estimate of the cross-spectrum, a first-order recursive smoothing can be applied [47]. While this approach might help to improve localization performance in scenarios with one target source, recursive smoothing reduces the ability of the localization algorithm to quickly respond to changes in source activity, which is particularly important for complex multi-source scenarios. Furthermore, the optimal smoothing constant might depend on a variety of different factors, such as the number of active sources, the level of noise and the reverberation time. Thus, exponential smoothing is not considered in this chapter.

One possibility of accumulating evidence about the location of sound sources is to average the GCC function across time frames [1, 57]. This approach, which will be

referred to as *AVG*, has the potential advantage that activity corresponding to multiple sound sources can be considered per frame. Regarding the GMM-based localization model, the probability of sound-source activity is averaged over all frames.

Alternatively, the most likely source location can be estimated on a frame-by-frame basis, and all resulting short-time estimates can be pooled into a histogram [1, 3, 50]. Assuming that each of the active sound sources is most dominant across a reasonable number of time frames, the histogram will approximate the probability density function, PDF, of the true location of all active sound sources [3]. In addition, variations of time-frequency-based, T-F, histograms, might be considered where competing sources with different spectral contributions can be separated [2]. However, this implies that *a priori* knowledge about the spectral content of active sources is available. Moreover, as this chapter focuses on the localization of multiple speakers that show activity in a similar frequency range, the frame-based histogram technique, denoted as *HIST*, will be considered.

As discussed in [58], deciding what number of bins is used for the histogram analysis is a difficult task. While a high histogram resolution might be beneficial in scenarios with moderate reverberation, a higher variance of the TDOA estimates due to strong reverberation and noise can cause the histogram to have bimodal peaks, which will be erroneously interpreted as two active sources. Thus, the choice is a trade-off between resolution and robustness. In accordance with [79], it has been decided to use 37 histogram bins to cover the azimuth range of $[-90, 90^\circ]$ in steps of 5° , where each individual bin is chosen to represent the time delay of the corresponding anechoic HRTF. To increase the resolution of the final azimuth estimate, exponential interpolation is applied to refine the maximum peak position of the histogram analysis [84].

Recently, a maximum likelihood, ML, framework for localization has been presented by Woodruff and Wang [78, 79], which jointly performs segregation and localization. Although small improvements were reported in comparison to the histogram approach [79], the computational complexity of the resulting search space is only feasible if the number of target sources is low, for instance three. Yet, because acoustic scenes with up to six competing speakers are used for evaluation in this chapter, the ML approach is not considered.

In order to address the problem of moving sources, other approaches aim at tracking the sound-source positions across time by employing *statistical particle filtering*, PF, techniques and *hidden Markov models*, HMMs [26, 61, 62, 76, 80]. But since the position of sound sources is assumed to be stationary throughout the time interval over which the localization information is integrated, these methods are not considered in this chapter. For the application of binaural analysis in combination with particle filtering, see [70], this volume.

The impact of temporal integration on sound-source localization is exemplified in Fig. 3 for a binaural mixture with three competing speakers in a reverberant environment with $T_{60} = 0.5$ s. More specifically, a comparison of two temporal integration strategies, namely, *averaging* versus *histogram*, is shown for two GCC-based methods, W_{CC} and W_{PHAT} weighting, and the GMM-based approach. In contrast to the conventional GCC-CC pattern shown in panel (a), the PHAT weighting in panel

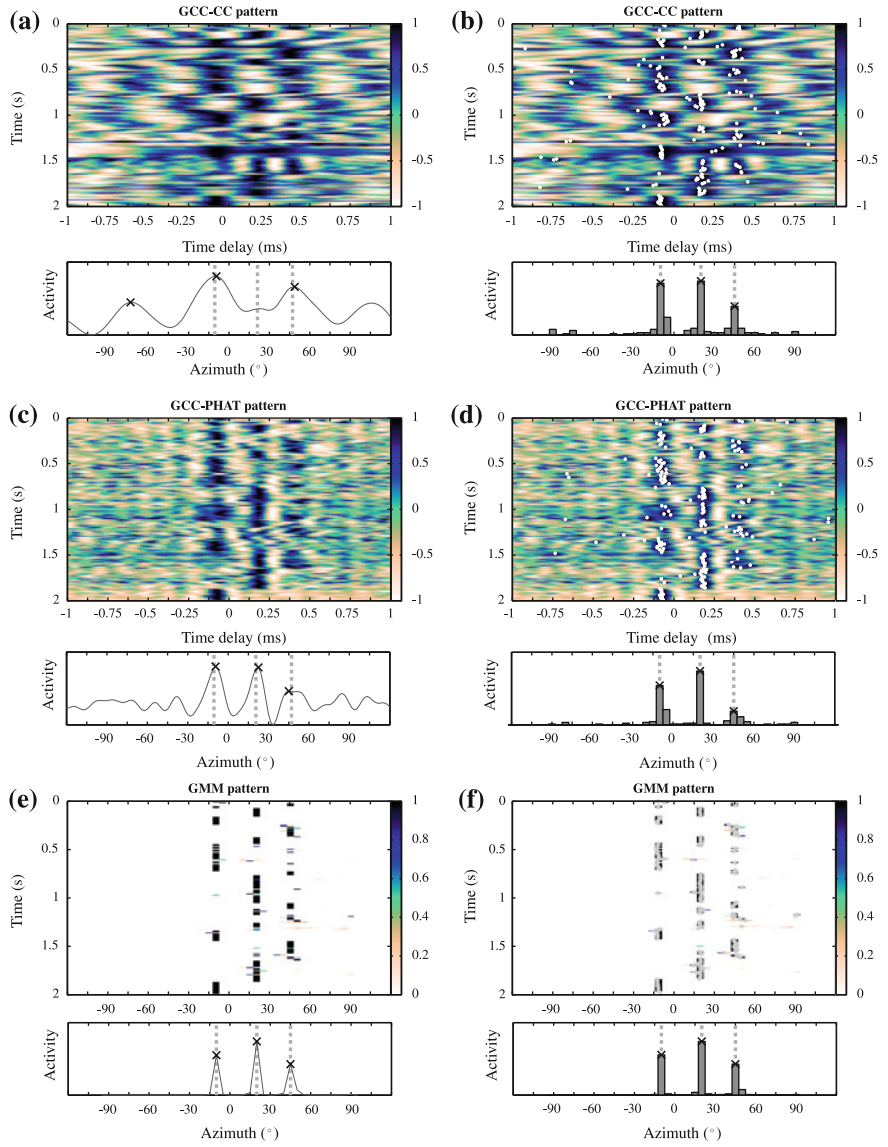


Fig. 3 Influence of two temporal integration strategies on localizing three competing talkers positioned at -10 , 20 and 35° in a reverberant room with $T_{60} = 0.5$ s. **(a, c)** Averaging of the GCC function across time. **(e)** GMM-based approach where the frame-based probability of sound-source direction is averaged over time. **(b, d, f)** Histogram-based integration. Dots represent the short-time localization estimates on a frame-by-frame basis. The estimated azimuth of the three speakers is marked by the *black crosses*, whereas their true position is indicated by *dashed vertical lines*

(c) produces sharper peaks, therefore, is able to resolve the positions of all three speakers. When using the histogram-based integration, both GCC–CC and GCC–PHAT achieve accurate predictions of the true speaker locations that are indicated by the vertical lines. The GMM-based approach shows the most prominent peaks of all methods at the true positions of the speakers for both integration strategies, where hardly any secondary peaks are visible.

To systematically compare the impact of these two temporal-integration strategies on localization performance, binaural mixtures of 2 s duration with up to six competing talkers are created, and the ability of various methods to predict the azimuth of all active speakers within $\pm 5^\circ$ accuracy is evaluated. The following acoustical parameters were varied,

- Number of competing speakers, ranging from one to six
- Randomized azimuth within $[-90, 90^\circ]$ with a minimum separation of 10°
- Simulated BRIRs ranging from $T_{60} = 0.2$ to $T_{60} = 1.05$ s

The experimental results are shown in Fig. 4 as a function of the reverberation time. Results are averaged over the number of competing speakers. In comparison to the frame-based localization accuracy reported in Sect. 4, the temporal integration significantly reduces the impact of reverberation on localization performance. In general, averaging the GCC pattern across time—dashed lines—is less robust than the histogram-based approach—solid lines—where short-time localization estimates are pooled across time. This is especially evident for the conventional GCC–CC method where the broad peaks in the accumulated GCC response prevent the detection of spatially close speakers—as seen in Fig. 3. Furthermore the averaging will integrate spurious peaks caused by reverberation, which might be erroneously considered as sound-source activity. In contrast, the histogram approach only considers the

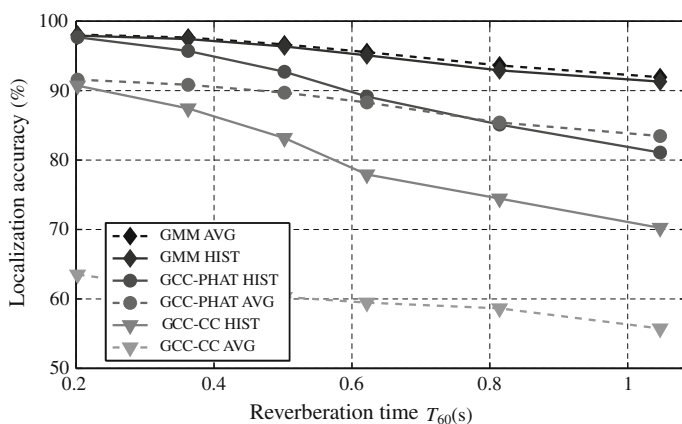


Fig. 4 Average performance of localizing up to six competing speakers with an accuracy of $\pm 5^\circ$ as a function of the reverberation time, T_{60} , for various approaches. The *dashed line* and the *solid line* indicate the two temporal integration strategies, namely, averaging and histogram-based integration

most salient source location on a frame-by-frame basis, thus focusing on the most reliable information. Consequently, the histogram-based integration of short-time localization estimates is very effective and considerably improves the robustness against the detrimental effect of reverberation. Regarding the GMM approach, a marginal benefit over the histogram-based integration is achieved when the probability of sound-source activity is averaged over time. This can be explained by the observation that the azimuth-dependent probability of sound-source activity is almost binary on a frame-by-frame basis—see Fig. 3—suggesting that each frame is approximately dominated by one individual sound source.

Overall, the PHAT weighting is substantially more robust than the conventional GCC–CC. Because the PHAT weighting already provides a sharp representation of the estimated time delay with strongly reduced secondary peaks—see Fig. 3—the additional improvement provided by the histogram integration is smaller than for the GCC–CC, most noticeably at short reverberation times. In anechoic conditions, GCC–PHATHIST performs as well as the GMM approach. But with increasing reverberation time, the multi-conditional training and the joint analysis of ITDs and ILDs enable the GMM-based localization method to be more robust in reverberant multi-source scenarios.

This benefit of the GMM-based approach over the GCC–PHATHIST system is presented in more detail in Fig. 5, where the localization performance is individually shown as a function of the number of competing talkers and the reverberation time. With an increasing number of speakers, the amount of reverberation has a stronger impact on the localization performance of the GCC–PHATHIST approach, as seen in panel (a). This dependency of the localization performance on the reverberation time is substantially reduced in panel (b), showing the robustness of the GMM-based approach.

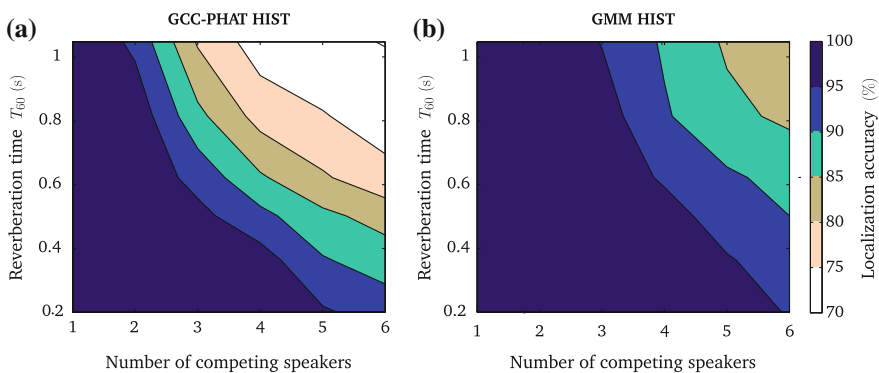


Fig. 5 Sound-source localization accuracy in % as a function of the number of competing speakers and the reverberation time for two approaches. **a** GCC–PHAT HIST. **b** GMM HIST

Table 1 Average localization accuracy in % for two sets of BRIRs. **a** Simulated BRIRs based on the KEMAR database [30]. **b** Measured BRIRs based on the HATS taken from the Surrey database [35]

BRIRs	Methods	# competing speakers						
		One	Two	Three	Four	Five	Six	Mean
(a) Simulated, KEMAR $T_{60} = \{0.36, 0.5, 0.62, 1.05\}$ s	GMM AVG	100	99.5	98.0	94.9	91.6	88.7	95.4
	GMM HIST	100	99.4	97.6	94.5	90.9	87.9	95.0
	GCC-PHAT HIST	100	97.8	92.2	86.7	82.1	79.3	89.7
	GCC-PHAT AVG	100	96.6	90.2	85.7	80.1	75.9	88.1
	GCC-CC HIST	96.8	85.2	79.5	74.3	72.1	70.3	79.7
	GCC-CC AVG	96.2	65.4	52.3	49.3	48.7	46.4	59.2
(b) Measured, HATS $T_{60} = \{0.32, 0.47, 0.68, 0.89\}$ s	GMM AVG	100	99.0	97.0	92.7	89.8	86.7	94.2
	GMM HIST	100	98.9	96.4	92.3	89.4	86.2	93.9
	GCC-PHAT HIST	99.9	98.3	95.7	89.7	84.8	80.7	91.5
	GCC-PHAT AVG	99.3	96.7	90.9	84.2	80.5	75.4	87.8
	GCC-CC HIST	93.9	82.7	76.4	71.0	69.1	67.3	76.7
	GCC-CC AVG	90.4	54.5	46.4	41.9	40.8	38.8	52.1

Generalization to Real Recordings

An important question is to what extent the results obtained with simulated BRIRs can be compared to recorded BRIRs. Therefore, the localization performance of simulated BRIRs is compared with a set of measured BRIRs. To allow for a fair comparison, a subset of the simulated BRIRs, $T_{60} = \{0.36, 0.5, 0.62 \text{ and } 1.05 \text{ s}\}$, was selected such that the reverberation times are as close as possible to the measured BRIRs, $T_{60} = \{0.32, 0.47, 0.68 \text{ and } 0.89 \text{ s}\}$ —see Sect. 2 for details. Furthermore, this comparison allows for assessing of how well the localization methods, which have all been trained on one particular artificial head, KEMAR, are able to generalize to the recorded BRIRs, which are based on a different artificial head, HATS.

The analysis involves binaural multi-source mixtures containing between one and six competing speakers that are created using both simulated and measured BRIRs. In Table 1, the localization accuracy of all tested methods is shown separately for (a), the simulated BRIRs based on the KEMAR artificial head and (b), the measured BRIRs based on the HATS artificial head. Results are averaged across all reverberation times. By comparing the mean values for different conditions, it can be seen that the overall performance for the measured BRIRs is fairly well reproduced by the set of simulated BRIRs. Thus, the differences in localization performance evaluated with simulated BRIRs are also valid for real life BRIRs. This is an important statement, justifying the usage of simulation tools for the development of localization algorithms. Furthermore, although the binaural-localization models are calibrated for one particular artificial head, localization performance does not degrade substantially when they are applied in the context of a different binaural-recording setup. Nevertheless, to minimize the sensitivity of the GMM-based localization model to a specific artificial head, the multi-conditional training stage can be readily adopted

to include various sets of different HRTFs. Alternatively, it is possible to employ generic head models if only the coarse characteristics of the human head should be captured [13, 28]. Moreover, supervised learning of binaural cues can also be applied in the field of robotics, which is discussed in [5], this volume.

6 Localization of Speakers in the Presence of Interfering Noise

When all active sound sources are assumed to be speakers, it is reasonable to cluster the localization information across time and to treat the most significant peaks as estimated source positions. However, if speech activity is corrupted by environmental noise, the task becomes much more difficult and a prominent peak might as well correspond to the position of a noise source. Therefore, a distinction between speech and noise sources is required in order to reliably select sound-source activity that originates from active speakers. In the following, the application of binaural cues to the problem of sound-source segregation is considered.

6.1 Segregation of Individual Sound Sources

In order to distinguish between speech and noise sources, the time-frequency, T-F, representation of multi-source mixtures will be segmented according to the estimated azimuth of sound sources. Assuming that sound sources are spatially separated, all T-F units that belong to one particular sound-source direction will be assumed to belong to the same acoustic source. This source segregation can subsequently be used to control a missing data classifier.

The GMM-based approach to binaural sound-source localization described in Sect. 3.3 was shown to accurately predict the location of up to six competing speakers in reverberation. Instead of using the most prominent peaks in the azimuth histogram as estimated sound-source positions, each local peak in the azimuth histogram will now be considered as a speech-source candidate. The corresponding histogram-bin indices are used to form a set of M candidate positions, $L = \{\ell_1, \dots, \ell_M\}$. Because the GMM-based approach extracts the likelihood of sound-source activity in individual frequency channels, the resulting spatial log-likelihood function, $\mathcal{L}(t, f, k)$, can be used to determine the contribution of all M candidate positions on a time-frequency, T-F, basis as

$$\mathcal{M}_m(t, f) = \begin{cases} 1 & \text{if } m = \arg \max_{k \in L} \mathcal{L}(t, f, k) \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

The resulting estimated *binary mask*, $\mathcal{M}_m(t, f)$, is a binary decision whether the m -th candidate has been the most dominant source in a particular T-F unit. The

binary mask has a wide variety of different application areas, among them automatic speech and speaker recognition [21, 51, 52] as well as speech enhancement [40]. Due to the promising results that were obtained with the *ideal binary mask*, IBM, where the optimal segregation is known *a priori*, the estimation of the ideal binary mask has been proposed as the main goal of computational auditory scene analysis [75].

Speech-Detection Module

In the following, it will be discussed how the estimated binary mask according to (8) can be used to select the most-likely speech sources among a set of candidate positions. The estimated binary mask can be used to perform *missing data*, MD, classification, where only a subset—indicated by the binary mask—of all time-frequency, T-F, units are evaluated by the classifier, namely those that are assumed to contain reliable information about the target source [21]. In this way, it is possible to selectively analyze and classify individual properties of one particular target source in the presence of other competing sources. Note that the concept of missing data is closely related to the auditory phenomenon of masking, where parts of the target source might be obscured and are, therefore, *missing* in the presence of other interfering sources [55, 75]. To distinguish between speech and noise sources, the amount of spectral fluctuation in individual Gammatone channels is a good descriptor that can be used to exploit the distinct spectral characteristic between speech and noise signals [49]. Based on a smoothed envelope, e_f , obtained by low-pass filtering the half-wave rectified output of the f th Gammatone channel with a time constant of 10 ms, the mean absolute deviation of the envelope over B samples is calculated as

$$\mathcal{F}(t, f) = \frac{1}{B} \sum_{i=0}^{B-1} |e_f(t \cdot B/2 - i) - \bar{e}_f|, \quad (9)$$

where \bar{e}_f reflects the mean envelope of the t -th frame. Note that the left and the right ear signals are averaged prior to envelope extraction. This feature, $\mathcal{F}(t, f)$, is subsequently modeled by two GMMs, denoted as λ_{Speech} and λ_{Noise} , reflecting the feature distribution for a large amount of randomly selected speech and noise files [49, 51]. Incorporating this *a priori* knowledge about the spectral characteristics of speech and noise signals can be viewed as an implementation of schema-driven processing. Given the estimated mask, \mathcal{M}_m , the two GMMs, λ_{Speech} and λ_{Noise} , and the extracted feature space, \mathcal{F} , the log-likelihood ratio of speech activity for the m -th candidate can be derived as

$$p_m = \log \left(\frac{p(\mathcal{F} | \lambda_{\text{Speech}}, \mathcal{M}_m)}{p(\mathcal{F} | \lambda_{\text{Noise}}, \mathcal{M}_m)} \right). \quad (10)$$

In order to emphasize speech-source candidates that are more frequently active in the acoustic scene, the log-likelihood ratio of speech activity is weighted with

the *a priori* probability of sound-sources activity, which is approximated by the normalized azimuth-histogram value of the corresponding candidate [49]. Although other weighting schemes can be considered, it was found that putting equal weight on the log-likelihood ratio obtained from the MD classifier and on the histogram-based localization information leads to good results. Finally, these weighted log-likelihood ratios of all M candidates are ranked in descending order, and the azimuth positions corresponding to the highest values are used to reflect the most likely speech source positions. In this way, the plain peak selection based on the most dominant localization information is supported by evidence about the source characteristic, being either speech-like or noise-like, therefore allowing for a distinction between speech and noise signals. The localization based on this *speech-detection module* will be referred to as *GMMSDM*.

Of course, other unique properties of speech signals might be considered at this stage as well, and a joint analysis of multiple complementary features is conceivable to further improve the ability to distinguish between speech and noise signals. As reported by [40, 42], the *amplitude-modulation spectrogram* is an effective feature that provides a reliable discrimination between speech and noise. Furthermore, it has been shown that also the distribution of reliable T-F units in the estimated binary mask, \mathcal{M}_m , contains information about the type of source, where the binary pattern shows a more compact representation for speech sources than for noise signals [48].

6.2 Influence of the Spatial Diffuseness of Interfering Noise

The impact of environmental noise on the ability to localize speakers does not only depend on the overall signal-to-noise ratio, but furthermore on its spatial distribution. Nevertheless, the vast majority of studies have investigated the influence of diffuse noise on sound-source localization [1, 16, 17, 79], which complies with the assumption of the GCC-based approach. However, the assumption of a diffuse noise field is not necessarily realistic for a real-life scenario. As recently analyzed by [58], real recordings of noise scenarios show a substantial amount of correlation, where the maximum value of the normalized cross-correlation function has been used as an indication of the amount of spatial correlation between the two microphones. Their experimental results showed that the conventional GCC method was superior to the PHAT weighting for acoustic conditions in which the noise had a high degree of correlation [58].

Therefore, the aim of this section is to investigate the impact of noise diffuseness on speaker-localization accuracy. More specifically, the influence of the noise characteristic is analyzed by systematically varying the amount of correlation of the noise between the left and the right ear signals. Therefore, different realizations of a particular noise type are filtered with BRIRs corresponding to a predefined number of randomly-selected azimuth directions. Note that for a given noise signal, each azimuth direction may be only selected once. By systematically varying the number of azimuth directions that contribute to the overall noise field from 1 to 37, the spatial

characteristic of the resulting noise can be gradually changed from a compact noise source located at one particular azimuth direction to a noise field where the energy is uniformly distributed across all 37 sound-source directions, thus approximating a diffuse noise field. The spatial diffuseness of the resulting noise field is specified by relating the number of noise realizations that contribute to the overall noise signal to the total number of discrete sound-source directions, ranging from $100 \cdot \frac{1}{37} = 2.7$ to $100 \cdot \frac{37}{37} = 100\%$.

In order to quantify the amount of correlation between the left and the right ear signals, the short-time coherence is estimated for 20 ms frames. The resulting coherence is averaged over time and shown in Fig. 6 as a function of frequency for noise signals consisting of 1, 3, 9, 19 and 37 superimposed realizations of randomly-selected azimuth directions. Whereas the average coherence in panel (a) is based on noise signals in anechoic conditions, panel (b) shows the additional influence of reverberation, namely, $T_{60} = 0.36$ s. It can be observed that the coherence functions systematically decrease with increasing number of noise realizations that contribute to the overall noise field. Furthermore, when comparing panel (a) and (b), it can be seen that in addition to the number of noise realizations, reverberation has a decorrelating effect, decreasing the correlation between the left and the right ear signals.

Now, the influence of interfering noise on localization performance is analyzed for binaural multi-talker mixtures with up to four competing talkers. Speech is corrupted with noise with the spatial distribution being gradually changed from compact noise to spatially diffuse noise. The following acoustic conditions are varied,

- Number of concurrent speakers ranging from one to four
- Number of interfering noise sources, 1, 3, 9, 19 and 37

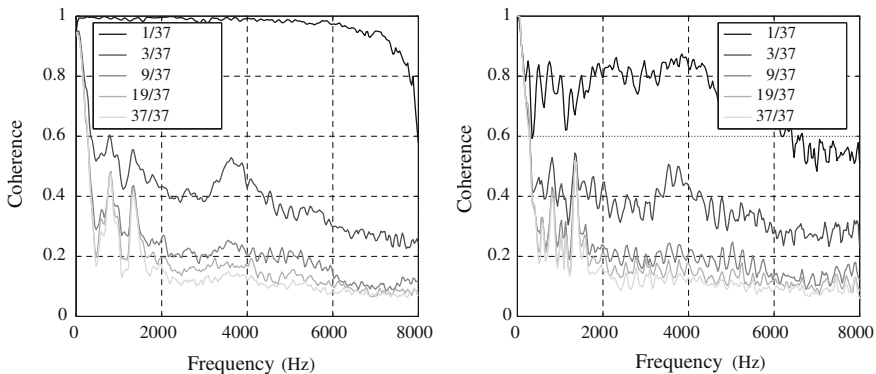


Fig. 6 Average short-time coherence estimates between the *left* and the *right* ear signals in response to various simulated noise signals. The individual noise signals consist of 1, 3, 9, 19 and 37 superimposed realizations of factory noise excerpts and are filtered with BRIRs corresponding to randomly selected azimuth directions. **a** Results for $T_{60} = 0$ s. **b** Results for $T_{60} = 0.36$ s. See Sect. 6.2 for details

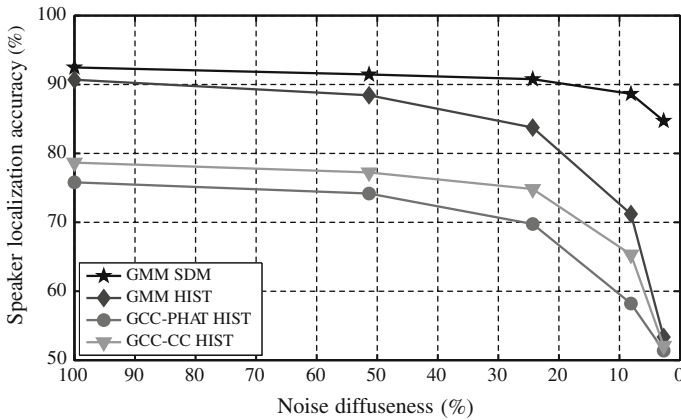


Fig. 7 Accuracy of localizing up to four competing speakers in a reverberant room, $T_{60} = 0.36$ s, as a function of the spatial diffuseness of the interfering noise. Performance is averaged over three SNRs, that is, 10, 5 and 0 dB, and three types of background noise, namely, factory, babble and speech-shaped noise

- Three noise types, namely, factory noise, babble noise and speech-shaped noise
- SNR between speech and noise, that is, 10, 5 and 0 dB

The performance of localizing up to four competing talkers within $\pm 5^\circ$ accuracy is presented in Fig. 7 as a function of the spatial diffuseness of the noise. Results are averaged over the number of competing talkers, the three noise types and the three SNRs. In general, the presence of noise imposes serious problems for the GCC-based approaches using either the W_{CC} or the W_{PHAT} weighting. In contrast to the results presented in Sect. 5, the PHAT weighting performs worse than the classical GCC-CC. This may be attributed to the whitening process, which equally weights all frequency components, thereby also amplifying the noise components. These results are in line with the observation of Perez-Lorenzo et al. [58], where the classical GCC-CC was reported to perform more robustly than the PHAT weighting for scenarios with correlated noise. Although GMMHIST appears to be more robust, the limiting factor that is shared by all of the aforementioned methods is that they solely exploit localization information. However, the most energetic components of speech are sparsely distributed in the presence of noise [20], thereby only a limited set of spectro-temporal units will be dominated by the sound-source direction of the speakers. Thus, as soon as the noise gets more directional, the noise energy is more compactly associated with a particular sound-source direction. As a result, the most dominant localization information will at a certain SNR inevitably correspond to the position of the interfering noise, which in turn reduces the overall speaker-localization accuracy. This observation corroborates the need for a distinction between speech and noise sources, especially for scenarios where the interfering noise has strong directional components. Such a distinction can be realized by using the speech-detection module described in Sect. 6.1, which effectively combines the localization analysis with a

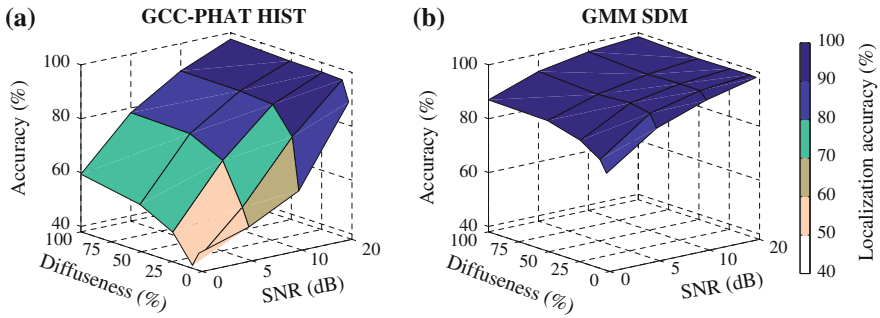


Fig. 8 Accuracy of localizing up to four competing speakers in a reverberant room, $T_{60} = 0.36$ s, as a function of the spatial diffuseness of the interfering noise and the signal-to-noise ratio

classification stage for selecting the most-likely speech sources according to (10). The experimental results shown in Fig. 7 demonstrate that the GMM SDM allows for a robust localization of up to four competing speakers, where the impact of directional noise is drastically reduced.

In Fig. 8, the localization performance of the two approaches GCC-PHAT HIST and GMMSDM is shown as a function of the SNR and the noise diffuseness. It can be seen that the performance of the PHAT approach systematically decreases with decreasing SNR, quite notably already at SNRs of 5–10 dB. Furthermore, the PHAT approach clearly suffers from interfering noise that is less diffuse, but correlated between the left and the right ears. In contrast, the GMMSDM approach that attempts to separate the contribution of individual sources on the basis of common spatial location in combination with employing a speech-detection module achieves robust localization performance over a wide range of experimental conditions. In summing up, it can be stated that interfering noise signal with a high degree of directional components will trigger the correlation-based approaches to localize the position of the noise components. Therefore, a distinction between speech and noise signals is required in order to enable a high speaker-localization accuracy in noisy environments.

7 Conclusions

This chapter presented an overview of binaural approaches to localizing multiple competing speakers in adverse acoustic scenarios. A fundamental limitation of many methods is that they assume single-path wave propagation, whereby performance inevitably decreases in the presence of reverberation and multiple competing sources. It was demonstrated that it is possible to incorporate the uncertainty of binaural cues in response to complex acoustic scenarios into a probabilistic model for robust sound-source localization, thus significantly improving the localization performance in the

presence of reverberation. To reliably estimate the location of multiple competing sound sources in reverberant environments, a histogram analysis of short-time localization estimates can substantially reduce the severe effect of reverberation. Furthermore, a comparison between simulated and recorded BRIRs has confirmed that the presented model produces accurate localization estimates for real-life scenarios and is able to generalize to an unseen artificial head, for which the system was not trained for. In general, considering both the impact of reverberation and noise imposes serious challenges for localization algorithms. A thorough analysis highlighted that in particular the spatial distribution of the noise field is a very important factor that strongly influences the performance of correlation-based localization algorithms, being most detrimental for GCC-based approaches if the interfering noise has a high degree of correlation between the left- and the right-ear signals. This problem can be overcome by separating the contribution of individual sound sources by means of estimating the binary mask. This binary mask can subsequently be used to control a missing data classifier, which is able to distinguish between sound-source activity emerging from speech and noise sources. It was shown that this joint analysis of localization information and source characteristic can be effectively used to achieve robust sound-source localization in very challenging acoustic scenarios.

Acknowledgments The authors are indebted to two anonymous reviewers for their constructive suggestions.

References

1. P. Aarabi. Self-localizing dynamic microphone arrays. *IEEE Trans. Sys., Man, Cybern., C*, 32(4):474–484, Nov. 2002.
2. P. Aarabi and S. Mavandadi. Robust sound localization using conditional time-frequency histograms. *Inf. Fusion*, 4(2):111–122, Sep. 2003.
3. P. Aarabi and S. Zaky. Iterative spatial probability based sound localization. In *Proceedings of the 4th World Multi-conference on Circuits, Systems, Computers and Communications*, Athens, Greece, Jul. 2000.
4. J. B. Allen and D. A. Berkley. Image method for efficiently simulating small-room acoustics. *J. Acoust. Soc. Am.*, 65(4):943–950, Apr. 1979.
5. S. Argentieri, A. Portello, M. Bernard, P. Danès, and B. Gas. Binaural systems in robotics. In J. Blauert, editor, *The technology of binaural listening*, chapter 9. Springer, Berlin-Heidelberg-New York NY, 2013.
6. R. Baumgartner, P. Majdak, and B. Laback. Assessment of sagittal-plane sound-localization performance in spatial-audio applications, chapter 4. In J. Blauert, editor, *The technology of binaural listening*. Springer–Berlin–Heidelberg–New York NY, 2013.
7. J. Benesty, J. Chen, and Y. Huang. Time-delay estimation via linear interpolation and cross correlation. *IEEE Trans. Speech Audio Process.* 12(5):509–519, 2004.
8. M. Bodden. Modeling human sound-source localization and the cocktail-party-effect. *Acta Acust./Acustica*, 1(1):43–55, 1993.
9. J. Braasch. Modelling of binaural hearing. In J. Blauert, editor, *Communication acoustics*, chapter 4, pages 75–108. Springer, Berlin, Germany, 2005.
10. A. S. Bregman. *Auditory scene analysis: The perceptual organization of sound*. The MIT Press, Cambridge, MA, USA, 1990.

11. A. W. Bronkhorst. The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acustica*, 86:117–128, 2000.
12. A. W. Bronkhorst and R. Plomp. Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing. *J. Acoust. Soc. Am.*, 92(6):3132–3139, Dec. 1992.
13. C. P. Brown and R. O. Duda. A structural model for binaural sound synthesis. *IEEE Trans. Speech Audio Process.*, 6(5):476–488, Sep. 1998.
14. G. J. Brown and M. Cooke. Computational auditory scene analysis. *Comput. Speech Lang.*, 8(4):297–336, Oct. 1994.
15. G. C. Carter, A. H. Nuttall, and P. G. Cable. The smoothed coherence transform. *Proceedings of the IEEE*, 61(10):1497–1498, Oct. 1973.
16. J. Chen, J. Benesty, and Y. Huang. Robust time delay estimation exploiting redundancy among multiple microphones. *IEEE Trans. Acoust., Speech, Signal Process.*, 11(6):549–557, 2003.
17. J. Chen, J. Benesty, and Y. A. Huang. Performance of GCC- and AMDF-based time-delay estimation in practical reverberant environments. *J. Appl. Signal Process.*, 1:25–36, 2005.
18. J. Chen, J. Benesty, and Y. A. Huang. Time delay estimation in room acoustic environments: An overview. *J. Appl. Signal Process.*, 2006:1–19, 2006.
19. E. C. Cherry. Some experiments on the recognition of speech, with one and two ears. *J. Acoust. Soc. Am.*, 25(5):975–979, Sep. 1953.
20. M. Cooke. A glimpsing model of speech perception in noise. *J. Acoust. Soc. Am.*, 199(3):1562–1573, Mar. 2006.
21. M. Cooke, P. Green, L. Josifovski, and A. Vizinho. Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Commun.*, 34:267–285, 2001.
22. M. Cooke and T.-W. Lee. Speech separation and recognition competition. URL <http://staffwww.dcs.shef.ac.uk/people/M.Cooke/SpeechSeparationChallenge.htm>, accessed on 15th January 2013, 2006.
23. C. J. Darwin. Auditory grouping. *Trends Cogn. Sci.*, 1(1):327–333, Dec. 1997.
24. M. S. Datum, F. Palmieri, and A. Moiseff. An artificial neural network for sound localization using binaural cues. *J. Acoust. Soc. Am.*, 100(1):372–383, Jul. 1996.
25. J. DiBiase, H. Silverman, and M. Brandstein. Robust localization in reverberant rooms. In M. Brandstein and D. Ward, editors, *Microphone arrays: Signal processing techniques and applications*, chapter 8, pages 157–180. Springer, Berlin, Germany, 2001.
26. M. Dietz, S. D. Ewert, and V. Hohmann. Auditory model based direction estimation of concurrent speakers from binaural signals. *Speech Commun.*, 53(5):592–605, 2011.
27. G. Doblinger. Localization and tracking of acoustical sources. In E. Haensler and G. Schmidt, editors, *Topics in acoustic echo and noise control*, chapter 4, pages 91–124. Springer, Berlin, Germany, 2006.
28. R. O. Duda and W. L. Martens. Range dependence of the response of a spherical head model. *J. Acoust. Soc. Am.*, 104(5):3048–3058, Nov. 1998.
29. C. Faller and J. Merimaa. Source localization in complex listening situations: Selection of binaural cues based on interaural coherence. *J. Acoust. Soc. Am.*, 116(5):3075–3089, Nov. 2004.
30. W. G. Gardner and K. D. Martin. HRTF measurements of a KEMAR dummy-head microphone. Technical report, # 280, MIT Media Lab, Perceptual Computing, Cambridge, MA, USA, 1994.
31. B. R. Glasberg and B. C. J. Moore. Derivation of auditory filter shapes from notched-noise data. *Hear. Res.*, 47(1–2):103–138, Aug. 1990.
32. T. Gustafsson, B. D. Rao, and M. Trivedi. Analysis of time-delay estimation in reverberant environments. In *Proc. ICASSP*, pages 2097–2100, Orlando, Florida, USA, May 2002.
33. S. Harding, J. Barker, and G. Brown. Mask estimation for missing data speech recognition based on statistics of binaural interaction. *IEEE Trans. Audio, Speech, Lang. Process.*, 14(1):58–67, Jan. 2006.
34. J.-S. Hu and W.-H. Liu. Location classification of nonstationary sound sources using binaural room distribution patterns. *IEEE Trans. Audio, Speech, Lang. Process.*, 17(4):682–692, May 2009.

35. C. Hummersone, R. Mason, and T. Brookes. Dynamic precedence effect modelling for source separation in reverberant environments. *IEEE Trans. Audio, Speech, Lang. Process.*, 18(7):1867–1871, Sep. 2010.
36. G. Jacovitti and G. Scarano. Discrete time techniques for time delay estimation. *IEEE Trans. Signal Process.*, 41(2):525–533, Feb. 1993.
37. M. Jeub, M. Schäfer, and P. Vary. A binaural room impulse response database for the evaluation of dereverberation algorithms. *Proc. Intl. Conf. Digital Signal Process. (DSP)*, pages 1–5, Jul. 2009.
38. A. Jourjine, S. Rickard, and Yilmaz. Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures. In *Proc. ICASSP*, pages 2985–2988, Istanbul, Turkey, Jun. 2000.
39. H. Kayser, S. D. Ewert, T. Rohdenburg, V. Hohmann, and B. Kollmeier. Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses. *EURASIP J. Adv. Sig. Proc.*, 2009.
40. G. Kim, Y. Lu, Y. Hu, and P. C. Loizou. An algorithm that improves speech intelligibility in noise for normal-hearing listeners. *J. Acoust. Soc. Am.*, 126(3):1486–1494, Sep. 2009.
41. C. H. Knapp and G. C. Carter. The generalized correlation method for estimation of time delay. *IEEE Trans. Acoust., Speech, Signal Process.*, ASSP-24(4):320–327, Aug. 1976.
42. B. Kollmeier and R. Koch. Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction. *J. Acoust. Soc. Am.*, 95(3):1593–1602, Mar. 1994.
43. E. H. A. Langendijk and A. W. Bronkhorst. Contribution of spectral cues to human sound localization. *J. Acoust. Soc. Am.*, 112(4):1583–1596, Oct. 2002.
44. W. Lindemann. Extension of a binaural cross-correlation model by contralateral inhibition. I. Simulation of lateralization for stationary signals. *J. Acoust. Soc. Am.*, 80(6):1608–1622, Dec. 1986.
45. W. Lindemann. Extension of a binaural cross-correlation model by contralateral inhibition. II. The law of the first wave front. *J. Acoust. Soc. Am.*, 80(6):1623–1630, Dec. 1986.
46. R. F. Lyon. A computational model of binaural localization and separation. In *Proc. ICASSP*, pages 1148–1151, Boston, Massachusetts, USA, Apr. 1983.
47. N. Madhu and R. Martin. Acoustic source localization with microphone arrays. In R. Martin, U. Heute, and C. Antweiler, editors, *Advances in Digital Speech Transmission*, chapter 6, pages 135–170. Wiley, 2008.
48. T. May and S. van de Par. Blind estimation of the number of speech sources in reverberant multisource scenarios based on binaural signals. in *Proc. IWAENC*, Aachen, Germany, Sep. 2012.
49. T. May, S. van de Par, and A. Kohlrausch. Binaural detection of speech sources in complex acoustic scenes. In *Proc. WASPAA*, pages 241–244, New Paltz, NY, USA, Oct. 2011.
50. T. May, S. van de Par, and A. Kohlrausch. A probabilistic model for robust localization based on a binaural auditory front-end. *IEEE Trans. Audio, Speech, Lang. Process.*, 19(1):1–13, Jan. 2011.
51. T. May, S. van de Par, and A. Kohlrausch. A binaural scene analyzer for joint localization and recognition of speakers in the presence of interfering noise sources and reverberation. *IEEE Trans. Audio, Speech, Lang. Process.*, 20(7):2016–2030, Sep. 2012.
52. T. May, S. van de Par, and A. Kohlrausch. Noise-robust speaker recognition combining missing data techniques and universal background modeling. *IEEE Trans. Audio, Speech, Lang. Process.*, 20(1):108–121, Jan. 2012.
53. R. Meddis, M. J. Hewitt, and T. M. Shackleton. Implementation details of a computation model of the inner hair-cell auditory-nerve synapse. *J. Acoust. Soc. Am.*, 87(4):1813–1816, Apr. 1990.
54. R. Meddis and E. A. Lopez-Poveda. Auditory periphery: From pinna to auditory nerve. In R. Meddis, E. A. Lopez-Poveda, R. R. Fay, and A. N. Popper, editors, *Computational models of the auditory system*, volume 35, chapter 2, pages 7–38. Springer, New York, 2010.
55. B. C. J. Moore. *An introduction to the psychology of hearing*. Academic Press, San Diego, California, USA, 5th edition, 2003.

56. J. Nix and V. Hohmann. Sound source localization in real sound fields based on empirical statistics of interaural parameters. *J. Acoust. Soc. Am.*, 119(1):463–479, Jan. 2006.
57. K. J. Palomäki, G. J. Brown, and D. L. Wang. A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation. *Speech Commun.*, 43(4):361–378, 2004.
58. J. Perez-Lorenzo, R. Viciano-Abad, P. Reche-Lopez, F. Rivas, and J. Escolano. Evaluation of generalized cross-correlation methods for direction of arrival estimation using two microphones in real environments. *Appl. Acoust.*, 73(8):698–712, Aug. 2012.
59. V. C. Raykar, B. Yegnanarayana, S. R. M. Prasanna, and R. Duraiswami. Speaker localization using excitation source information in speech. *IEEE Trans. Speech Audio Process.*, 13(5):751–761, Sep. 2005.
60. L. Rayleigh. On our perception of sound direction. *Philos. Mag.*, 13:214–232, 1907.
61. N. Roman and D. L. Wang. Binaural tracking of multiple moving sources. In *Proc. ICASSP*, volume 5, pages 149–152, Hong Kong, China, Apr. 2003.
62. N. Roman and D. L. Wang. Binaural tracking of multiple moving sources. *IEEE Trans. Audio, Speech, Lang. Process.*, 16(4):728–739, 2008.
63. N. Roman, D. L. Wang, and G. J. Brown. Speech segregation based on sound localization. *J. Acoust. Soc. Am.*, 114(4):2236–2252, Oct. 2003.
64. R. Roy and T. Kailath. ESPRIT - estimation of signal parameters via rotational invariance techniques. *IEEE Trans. Acoust., Speech, Signal Process.*, 37(7):984–995, Jul. 1989.
65. S. M. Schimmel, M. F. Müller, and N. Dillier. A fast and accurate “shoebox” room acoustics simulator. In *Proc. ICASSP*, pages 241–244, Taipei, Taiwan, Apr. 2009.
66. R. O. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Trans. Antennas Propagat.*, AP-34(3):276–280, Mar. 1986.
67. M. R. Schroeder. New method for measuring reverberation time. *J. Acoust. Soc. Am.*, 37(3):409–412, 1965.
68. C. L. Searle, L. D. Braida, D. R. Cuddy, and M. F. Davis. Binaural pinna disparity: another auditory localization cue. *J. Acoust. Soc. Am.*, 57(2):448–455, Feb. 1975.
69. T. M. Shackleton, R. Meddis, and M. J. Hewitt. Across frequency integration in a model of lateralization. *J. Acoust. Soc. Am.*, 91(4):2276–2279, Apr. 1992.
70. C. Spille, B. Meyer, M. Dietz, and V. Hohmann. Binaural scene analysis with multi-dimensional statistical filters, chapter 6. In J. Blauert, editor, *The technology of binaural listening*. Springer, Berlin-Heidelberg-New York NY, 2013.
71. R. M. Stern, A. S. Zeiberg, and C. Trahiotis. Lateralization of complex binaural stimuli: A weighted-image model. *J. Acoust. Soc. Am.*, 84(1):156–165, Jul. 1988.
72. C. J. Sumner, E. A. Lopez-Poveda, L. P. O’Mard, and R. Meddis. A revised model of the inner-hair cell and auditory-nerve complex. *J. Acoust. Soc. Am.*, 111(5):2178–2188, May 2002.
73. S. Tervo and T. Lokki. Interpolation methods for the SRP-PHAT algorithm. In *Proc. IWAENC*, Seattle, Washington, USA, Sep. 2008.
74. A. P. Varga, H. J. M. Steeneken, M. Tomlinson, and D. Jones. The NOISEX-92 study on the effect of additive noise on automatic speaker recognition. Technical report, Speech Research Unit, Defence Research Agency, Malvern, UK, 1992.
75. D. L. Wang and G. Brown, editors. *Computational auditory scene analysis: Principles, algorithms and applications*. John Wiley & Sons, Hoboken, NJ, USA, 2006.
76. D. B. Ward, E. A. Lehmann, and R. C. Williamson. Particle filtering algorithms for tracking an acoustic source in a reverberant environment. *IEEE Trans. Speech Audio Process.*, 11(6):826–836, Nov. 2003.
77. V. Willert, J. Eggert, J. Adamy, R. Stahl, and E. Körner. A probabilistic model for binaural sound localization. *IEEE Trans. Sys., Man, Cybern.*, B, 36(5):982–994, Oct. 2006.
78. J. Woodruff and D. L. Wang. Sequential organization of speech in reverberant environments by integrating monaural grouping and binaural localization. *IEEE Trans. Audio, Speech, Lang. Process.*, 18(7):1856–1866, Sep. 2010.
79. J. Woodruff and D. L. Wang. Binaural localization of multiple sources in reverberant and noisy environments. *IEEE Trans. Audio, Speech, Lang. Process.*, 20(5):1503–1512, Jul. 2012.

80. J. Woodruff and D. L. Wang. Binaural detection, localization, and segregation in reverberant environments based on joint pitch and azimuth cues. *IEEE Trans. Audio, Speech, Lang. Process.*, 21(4):806–815, Apr. 2013.
81. O. Yilmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Signal Process. Lett.*, 52(7):1830–1847, Jul. 2004.
82. P. Zakarauskas and M. S. Cynader. A computational theory of spectral cue localization. *J. Acoust. Soc. Am.*, 94(3):1323–1331, Sep. 1993.
83. C. Zhang, D. Florêncio, and Z. Zhang. Why does PHAT work well in low noise, reverberative environments? In *Proc. ICASSP*, pages 2565–2568, 2008.
84. L. Zhang and X. Wu. On cross correlation based discrete time delay estimation. In *Proc. ICASSP*, volume 4, pages 981–984, Philadelphia, Pennsylvania, USA, 2005.