A Probabilistic Model for Robust Localization Based on a Binaural Auditory Front-End

Tobias May, Steven van de Par, and Armin Kohlrausch

Abstract—**Although extensive research has been done in the field of machine-based localization, the degrading effect of reverberation and the presence of multiple sources on localization performance has remained a major problem. Motivated by the ability of the human auditory system to robustly analyze complex acoustic scenes, the associated peripheral stage is used in this paper as a front-end to estimate the azimuth of sound sources based on binaural signals. One classical approach to localize an acoustic source in the horizontal plane is to estimate the interaural time difference (ITD) between both ears by searching for the maximum in the cross-correlation function. Apart from ITDs, the interaural level difference (ILD) can contribute to localization, especially at higher frequencies where the wavelength becomes smaller than the diameter of the head, leading to ambiguous ITD information. The interdependency of ITD and ILD on azimuth is a complex pattern that depends also on the room acoustics, and is therefore learned by azimuth-dependent Gaussian mixture models (GMMs). Multiconditional training is performed to take into account the variability of the binaural features which results from multiple sources and the effect of reverberation. The proposed localization model outperforms state-of-the-art localization techniques in simulated adverse acoustic conditions.**

Index Terms—**Localization, binaural, reverberation, auditory scene analysis (ASA), interaural time difference (ITD), interaural level difference (ILD).**

I. Introduction

The ability to localize sound sources in adverse acoustic environments is necessary for a wide range of applications, e.g., communication devices and hearing aids. Although extensive research has been done in the field of localization, the localization of multiple sources in adverse acoustic conditions has remained a challenging task. The performance of microphone array localization depends on the array configuration and generally increases with the number of microphones [1]. In contrast to microphone-array-based techniques, the performance of the human auditory system is very robust against the presence of multiple competing sources (cocktail party scenario) for tasks related to localization [2] and speech recognition [3], despite only exploring the acoustic mixture arriving at both ears. These remarkable capabilities of the human auditory system imply that in principle it is possible to analyze complex acoustic scenes and to independently localize and identify a desired source within an acoustic mixture based on binaural signals. This human ability to analyze complex acoustic mixtures is referred to as auditory scene analysis (ASA) [4]. One influential view on ASA is that the underlying mechanisms to retrieve information about a specific sound source can be divided into two processes: first, the human auditory system parses the acoustic scene into fragments (regions in the time-frequency plane), and second, fragments which may belong to the same acoustic object are grouped together. Because of this ability of the human auditory system to segment, group and integrate information of multiple sources in adverse acoustic conditions, research dealing with models of computational auditory scene analysis (CASA) is a growing field. A comprehensive overview of CASA and its relevance for applications in the field of automatic speaker recognition (ASR) and speech segregation can be found in [5]. The reliable estimation of the source position based on binaural signals is relevant not only for CASA systems, e.g., as applied in the area of robust speech recognition [6], [7], but further more is required for hearing aids and systems related to wearable augmented reality audio (WARA) [8]. Inspired by the robustness of the human auditory system, several studies have incorporated stages of human auditory processing to improve sound source localization in adverse acoustic conditions [9]–[11].

The two major cues which are exploited by the human auditory system to localize acoustic sources are interaural time and level differences. One of the classical approaches to measure the interaural time difference is to search for the main peak in the generalized cross-correlation (GCC) function [12], which estimates the interaural time difference (ITD) between the left and the right ears. The mathematical operation of the uniformly-weighted GCC between binaural signals is equivalent to the coincidence model suggested first in 1948 for describing human sound source localization [13]. Since then, several modifications and extensions have been proposed to explain the results obtained from psychoacoustic experiments. Another important cue exploited by the human auditory system is the interaural level difference (ILD), which is attributed to the head shadowing effects. The ILD was taken into account by incorporating the mechanisms of contralateral and temporal inhibition into the cross-correlation model [14]. In this way, several psychoacoustic phenomena related to the precedence effect could be successfully predicted [15]. A comprehensive review of the recent development of binaural models can be found in [16].

As indicated, the ILD cue plays an important role in localization, especially at higher frequencies where the wavelength

Manuscript received May 04, 2009; revised October 19, 2009; accepted January 12, 2010. Date of publication February 05, 2010; date of current version October 01, 2010. This work was supported by Philips Research. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Dan Ellis.

T. May and S. van de Par are with the Institute of Physics, University of Oldenburg, 26111 Oldenburg, Germany (e-mail: tobias.may@uni-oldenburg.de).

A. Kohlrausch is with Philips Research, NL-5656 AE Eindhoven, The Netherlands, and also with the Department of Human Technology Interaction, Eindhoven University of Technology, NL-5600 MB Eindhoven, The Netherlands.

Digital Object Identifier 10.1109/TASL.2010.2042128

1558-7916/$26.00 © 2010 IEEE

becomes smaller than the diameter of the head, leading to ambiguous ITD information. Nevertheless, there have been only a few attempts to combine both cues in a model for binaural sound localization. For example, the peak selection in the cross-correlation analysis was steered by the ILD cue in order to select the correct peak at higher frequencies where the cross-correlation function becomes ambiguous [17]. Nonetheless, the presence of reverberation has a stronger impact on the ILD cue than on the ITD [18], thus making such a peak selection procedure less reliable in adverse acoustic conditions. In addition, the dependence of ITD and ILD on azimuth is a complex, multimodal pattern that also depends on the reverberation and the presence of competing sources, and accordingly it can be best exploited by using a probabilistic model.

A combined evaluation of binaural cues has been successfully applied as a front-end for sound source segregation [19], where a joint feature space consisting of ITDs and ILDs was trained in ideal acoustic conditions in order to segregate a target source from interfering sources at different azimuth positions. The modeling was performed by an adaptive kernel density method, because the use of Gaussian mixture models (GMMs) had been reported to lead to issues related to the initialization process and to the problem of selecting the number of Gaussian components [19].

In [20] and [7], the effect of reverberation was included in a probabilistic model to predict a missing data mask for speaker recognition. A histogram technique was utilized to model the probability density function (pdf) of the binaural cues associated with the target source located at 0° azimuth. Although the performance in reverberation was reported to be comparable to the system established under anechoic acoustic conditions [19], the performance was sensitive to the source/receiver configuration which was used to train the model for the recognition of the binaural cues [20]. In [21], pdfs of interaural cues based on the fast Fourier transform (FFT), namely interaural phase differences (IPDs) and ILDs were measured by histograms in order to perform localization in nonstationary noise conditions. Those cues were integrated by combining their probabilities across frequency.

In this paper, a sound source localization model is presented that is robust against the presence of multiple sources, changes in source positions, and the impact of reverberation. Based on an auditory front-end, the complex interaction of ITDs and ILDs is learned by a probabilistic model that is trained under various acoustic conditions to obtain robustness. For single sound source localization, long analysis windows between 100–200 ms length are commonly applied to increase the robustness of localization in reverberation [22], [23]. To be able to resolve the time-dependent azimuth position of the most salient source in complex multi-source mixtures, which is required for many relevant applications (e.g., beamforming, tracking and CASA systems), a relatively short analysis window of 20 ms is used in this study. GMMs are used to learn the azimuth-dependent distribution of the binaural feature space consisting of both ITDs and ILDs. This feature space, based on time-domain analysis, is then compared to an FFT-based feature space consisting of IPD and ILD as described in [21]. Furthermore, the selection procedure for the number of Gaussian components will also be addressed. The straightforward approach is a manual selection by visual inspection, which will be compared to an unsupervised learning of Gaussian mixtures [24], where the model complexity is automatically determined. The performance of the GMM-based localization method is evaluated and compared with some state-of-the-art localization techniques in multi-source reverberant conditions. Finally, the ability of the model to generalize to unknown source/receiver configurations is discussed.

The paper is organized as follows. The next section describes the extraction of the binaural cues. Section III explains the details of the probabilistic model for sound source localization. In Section IV, the evaluation procedure is described and the performance in simulated echoic multisource scenarios is presented in Section V. A summary of the main findings and concluding remarks will be given in Section VI.

II. Binaural Cue Extraction

The two main cues enabling human sound source localization in the horizontal plane are interaural time and level differences, ITDs and ILDs, respectively. The ITD cue is most robust at low frequencies, whereas the ILD cue is predominantly used at higher frequencies [25]. The direction-dependent spectral modifications largely associated with the complex pinna shape are especially important for the elevation detection and to resolve front/back ambiguities. Because the localization task in this work is restricted to the frontal horizontal plane (zero elevation), the ability to discriminate sources in the vertical domain will not be investigated.

A. Auditory Front-End

The peripheral processing of the human auditory system is simulated by an auditory front-end consisting of a gammatone filterbank followed by inner hair cell-processing. This front-end is adopted from [19]. In order to resemble the frequency selectivity of the human cochlea, the signals arriving at the left and the right ear are decomposed into $N = 32$ auditory channels using a fourth-order gammatone filterbank. More specifically, phase-compensated gammatone filters are used to synchronize the binaural cues across auditory channels at a common time instance [26]. The gammatone channel responses are aligned by compensating for the group delays of the gammatone filters at their nominal center frequencies. According to [27], the channel center frequencies are equally distributed on the equivalent rectangular bandwidth (ERB) scale between 80 Hz and 5 kHz. Similar to the model described in [19], channel-dependent gains are applied to simulate the middle-ear transfer function, as determined by [28]. The neural transduction process in the inner hair cells is approximated by halfwave-rectification followed by a square-root compression. The resulting binaural auditory signals of the ith gammatone channel are represented by l_i and r_i respectively. Binaural cues are estimated using a rectangular window of 20 ms at a sampling frequency of $f_s = 44.1$ kHz (corresponding to $W = 882$ samples). An overlap between successive frames of 50% was applied, corresponding to a frame shift of 10 ms. This high temporal resolution was chosen to capture rapid changes in multi-source scenarios.

B. ITD

The time difference between the binaural auditory signals in the ith channel is estimated using the normalized cross-correlation function, which is defined in (1), shown at the bottom of the page, as a function of time lag τ and the frame number t. $\bar{l}_i$ and $\bar{r}_i$ denote the mean values of the left and right auditory signals and these are estimated over the frame number t. The cross-correlation function is evaluated for time lags within the range of $[-1, 1]$ ms, and its maximum corresponds to the estimated ITD (in samples)

$$\hat{\tau}_i(t) = \arg\max_{\tau} C_i(t, \tau). \tag{2}$$

The resolution of the ITD cue is limited by the sampling interval, whereas the actual time delay can lie between two successive samples. In order to increase the ITD accuracy while keeping the computational complexity moderate, exponential interpolation can be applied, as defined in (3) at the bottom of the page, around the estimated maximum $\hat{\tau}_i(t)$ of the cross-correlation function [29]. The fractional part $\hat{\delta}_i(t)$ can be considered to describe the interpolated peak position relative to the estimated integer peak position $\hat{\tau}_i(t)$, and the overall ITD estimate $\hat{\mathrm{itd}}_i(t)$ is then given in seconds by the combination of the integer and the fractional estimate

$$\hat{\mathrm{itd}}_i(t) = \left(\hat{\tau}_i\,(t) + \hat{\delta}_i\,(t)\right) / f_s. \tag{4}$$

In addition to the exponential interpolation, another classical approach was also tested, which describes the peak of the band-limited cross-correlation function by a parabola [30]. The performance of both interpolation methods were almost identical in low gammatone channels up to 1.5 kHz, whereas the exponential interpolation gave better results at higher frequencies and was therefore selected in the current study.

C. ILD

The interaural level difference is estimated by comparing the energy integrated across the time interval W between the left and right ears. The ILD cue in the ith gammatone channel expressed in dB is given by

$$\hat{\mathrm{ild}}_i(t) = 20\log_{10}\left(\frac{\sum_{n=0}^{W-1} r_i(t\cdot W/2 - n)^2}{\sum_{n=0}^{W-1} l_i(t\cdot W/2 - n)^2}\right). \tag{5}$$

Note that in (5), 20 instead of 10 is used to compensate for the square-root compression of the neural transduction process. A sound source positioned at the left-hand side will result in a negative ILD, whereas a positive ILD will be caused by a source lateralized to the right-hand side.

III. Model Architecture

A probabilistic model is used to estimate the position of a sound source from the set of binaural cues described in Section II. Therefore, GMMs are trained to recognize the azimuth-dependent pattern of the binaural cues. The training and the architecture of the model is described in the following sections.

A. Multi-Conditional Training

To achieve a robust localization performance, the model is trained in various simulated acoustic conditions to account for the variability of the binaural features caused by multiple sources and the effect of additional reverberation. As analyzed by [19], the distribution of binaural cues is dependent on the presence of an interfering source and its strength relative to the target source. To incorporate this effect into the model, the training sequences consist of a target source within the azimuth range of $[-50^\circ, 50^\circ]$ with an interfering source positioned at $\pm5^\circ, \pm10^\circ, \pm20^\circ, \pm30^\circ$, and $\pm40^\circ$ relative to the target azimuth (see Fig. 2). All target-/interfering-source combinations were presented at three different global signal-to-noise ratios (SNRs) of 20, 10, and 0 dB, as defined prior to spatialization. Moreover, the uncertainty of the binaural features attributed to the room reverberation is taken into account by using binaural room impulse responses (BRIRs). This approach is similar to the training procedure described in [7] and [20], where mixtures of multiple sources (target plus interfering source) were used in reverberation to obtain a more reliable model for identifying time-frequency elements (mask), which are associated with the target source only. This mask was used in the context of missing data speech recognition, where the recognition stage is based on the reliable components, while excluding time-frequency elements which are dominated by the interfering source. The authors showed that their model is robust against changes in the simulated room absorption which were not considered in the training stage. However, their system was sensitive to the relative placement of the source and the receiver within the room (source/receiver configuration) [20]. To improve the model performance in this respect, the training data in the current study were created using multiple source/receiver

$$C_i(t,\tau) = \frac{\sum_{n=0}^{W-1}(l_i(t\cdot W/2 - n) - \bar{l}_i)(r_i(t\cdot W/2 - n - \tau) - \bar{r}_i)}{\sqrt{\sum_{n=0}^{W-1}(l_i(t\cdot W/2 - n) - \bar{l}_i)^2}\sqrt{\sum_{n=0}^{W-1}(r_i(t\cdot W/2 - n - \tau) - \bar{r}_i)^2}} \tag{1}$$

$$\hat{\delta}_i(t) = \frac{\log C_i(t, \hat{\tau}_i(t)+1) - \log C_i(t, \hat{\tau}_i(t)-1)}{4\log C_i(t, \hat{\tau}_i(t)) - 2\log C_i(t, \hat{\tau}_i(t)-1) - 2\log C_i(t, \hat{\tau}_i(t)+1)} \tag{3}$$

configurations, where the BRIRs were synthesized by using the room simulation package developed by [31]. This software package combines a database of head-related transfer functions (HRTFs) [32] measured in anechoic conditions, with room reflections simulated according to the image source model [33]. To create the target and the interfering source signals, utterances of male speakers which were randomly selected from the TIMIT database [34] were convolved with the simulated BRIRs. Throughout the multiconditional training phase, the frequency-dependent absorption coefficients of the room were chosen to yield a constant reverberation time $T_{60} = 0.5$ s, in order to introduce the same amount of uncertainty for all gammatone channels. Note that only one level of reverberation was used to train the localization model. To obtain a reliable estimate of the target position, it is essential that the probabilistic model is trained only with binaural features which are associated with the target source. Thus, four criteria were employed to select the frames where the binaural features are dominated by the target source. Note that the last three criteria are monitored in all gammatone channels independently, whereas the first criterion is based on the signal prior to gammatone analysis. Firstly, an energy-based voice activity detector (VAD) was used to monitor the activity of the target source, and, a frame was considered to be silent and excluded if the energy level drops by more than 40 dB below the global maximum. Second, frames were considered for training only if the target source was stronger than the interfering source. This analysis compared the energy of the target source to the energy of the interfering source after spatialization. The signals of the left and the right ear were added prior to energy computation. Third, frames were removed when the height of the primary peak in the cross-correlation function was less then a threshold θ_c, assuming that the associated binaural cues are dominated by the room reflections. This third criterion was motivated by the fact that the amplitude of the cross-correlation reveals information about the ratio between the direct sound and the room reflections, which becomes low when the signals at the left and the right ear are dominated by reflections. The threshold was set to $\theta_c = 0.3$ by inspection, which still considers frames with low correlation between the binaural signals to incorporate the uncertainty of binaural cues resulting from adverse acoustic conditions into the training procedure. The fourth criterion removed frames from the training set, if the maximum of the cross-correlation function corresponded to one of the most lateral time lags ($\tau = \pm 44$).[1] For those time lags, it is assumed that the corresponding ITD of $[-1, 1]$ ms is outside the plausible range for the human head. Based on these four criteria, about 50% of the frames were removed.

B. Binaural Feature Space

As already pointed out, the ITD and the ILD cues contain complementary information about the source position, and can therefore be combined in a two-dimensional binaural feature space

$$\begin{aligned} X_i &= \{\vec{x}_{i,1}, \ldots, \vec{x}_{i,T}\} \\ &= \{(\hat{\mathrm{itd}}_i(1), \hat{\mathrm{ild}}_i(1)), \ldots, (\hat{\mathrm{itd}}_i(T), \hat{\mathrm{ild}}_i(T))\} \end{aligned} \qquad (6)$$

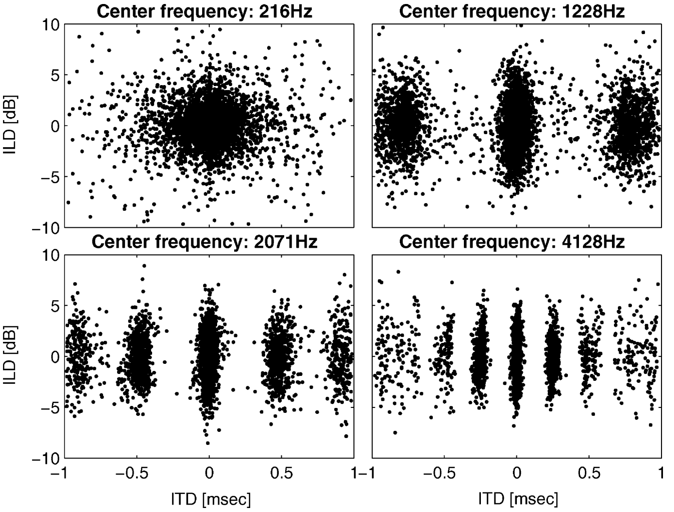

Fig. 1. Binaural feature space computed for a target source at 0° azimuth at different gammatone filter center frequencies under reverberation condition ($T_{60} = 0.5$ s). The number of clusters increases along the ITD feature dimension due to the ambiguous nature of the cross-correlation function at high frequencies.

where T represents the number of observations for gammatone channel i. This joint feature space of ITDs and ILDs is shown in Fig. 1 for a speech source at 0° azimuth using the multiconditional training. Each dot represents an observation of the binaural feature space for a single frame within a specific gammatone channel. The receiver (KEMAR head) was placed in the middle of the room, whereas the source was positioned at a radial distance of 1.5 m with respect to the receiver. The binaural cues were simulated in a room measuring $5.1 \times 7.1 \times 3$ m with a reverberation time of $T_{60} = 0.5$ s.

It can be observed in Fig. 1 that the interdependency of ITDs and ILDs results in complex patterns. At higher frequencies, where the wavelength is smaller than the diameter of the head, the ITD information becomes ambiguous. This effect is reflected by the number of distinct clusters in the binaural feature space, which systematically increases with the gammatone center frequency. The spread of the clusters can be related to the reverberation and the presence of an interfering source. Considering a target source at 0° azimuth in anechoic conditions without an interfering source, the distribution of ITDs and ILDs would be very narrow and hardly any side peaks would be observed.

To estimate the position of a sound source from a set of binaural cues, the complex pattern of the binaural feature space is learned by a probabilistic model. In [20], the pdf of the binaural feature space depending on the sound source azimuth was modeled by a histogram technique. In that study, two histograms were computed: one analyzing the binaural feature space for both target and interfering sources, and the second for the observations related to the target source only. The relation between these two histograms was used to derive the probability of a region which is dominated by the target source. The bin size of the histogram is the result of a tradeoff between the pdf resolution and the amount of data required for a sufficient training of the model. Furthermore, a threshold needs to be set for the histogram in order to control the potential effect of insufficient training on the pdf, which may occur for certain binaural

[1] Valid for a sampling frequency of 44.1 kHz.

feature combinations. Hence, ITD-ILD combinations were removed from the histogram if the number of counts was below a certain threshold, producing better estimates. It was also reported in [20] that the performance was sensitive to the histogram threshold and, more importantly, to the source/receiver configuration used for the simulation of the training data.

In order to overcome these limitations, Gaussian mixtures are chosen in the current study to model the probability density of the binaural cues for all azimuth positions. Because the binaural features tend to cluster in the feature space, the azimuth-dependent pdf can be modeled by the sum of superimposed Gaussian components. Also, the use of GMMs results in a smoother decision area than the histogram technique, which is expected to reduce the sensitivity of the model to unknown source/receiver configurations.

C. Gaussian Mixture Modeling

Gaussian mixture models are used to describe the direction-dependent distribution of the binaural feature space. Considering one specific sound source direction λ, a Gaussian mixture density for a D-dimensional feature vector $\vec{x}$ is the weighted sum of K Gaussian components [35]

$$p(\vec{x}\,|\,\lambda) = \sum_{j=1}^{K} w_j p_j(\vec{x}) \tag{7}$$

where $\vec{x}$ corresponds to the output of one specific gammatone channel. Each mixture component j is characterized by the component weight w_j, its mean vector $\vec{\mu_j}$ and the covariance matrix Σ_j. The variable λ represents the sound source properties. Furthermore, the mixture weights w_j satisfy $\sum_{j=1}^{K} w_j = 1$. Each of the K components is a D-variate Gaussian function given by

$$p_j(\vec{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma_j|^{1/2}} \exp\left[-\frac{1}{2}(\vec{x}-\vec{\mu_j})^t \Sigma_j^{-1}(\vec{x}-\vec{\mu_j})\right]. \tag{8}$$

The parameters required to uniquely describe the GMM can be summarized by the following notation:

$$\lambda = (w_j, \vec{\mu}_j, \Sigma_j) \quad \forall\ j = 1, \ldots, K. \tag{9}$$

Diagonal covariance matrices are used to describe the dependency between the ITD and the ILD features, since the clusters are orientated perpendicularly with respect to the ILD dimension (see Fig. 1). Moreover, the correlation between feature vector elements can be modeled, in principle, by a larger number of diagonal covariance matrices than the full covariance matrices, which are computationally more expensive [35].

Let $X = \{\vec{x}_1, \ldots, \vec{x}_T\}$ be a set of T observations of the D-dimensional feature vector $\vec{x}$. The log likelihood of the sound source λ for a single observation can be computed as follows:

$$\log p(\vec{x}_t\,|\,\lambda) = \log \sum_{j=1}^{K} w_j\, p_j(\vec{x}_t). \tag{10}$$

One GMM was chosen to model the pattern of the binaural feature space within each gammatone channel i for each sound source direction φ independently. Extending the log likelihood computation to N gammatone channels indicated by the index i and to S equally likely sound source directions represented by $\{\lambda_{i,\varphi_1}, \lambda_{i,\varphi_2}, \ldots, \lambda_{i,\varphi_S}\}$, the estimated sound source location is found by maximizing the log likelihood of the current observation $\vec{x}_{i,t}$

$$\hat{\mathrm{S}}(\vec{x}_{i,t}) = \underset{1\le k\le S}{\arg\max} \underbrace{\sum_{i=1}^{N} \underbrace{\log p(\vec{x}_{i,t}|\lambda_{i,\varphi_k})}_{\text{log likelihood}}}_{\text{across frequency integration}} . \tag{11}$$

This azimuth decision is made on a frame-by-frame basis in order to capture the time-dependent characteristics of multiple acoustic sources. In contrast to summing the binaural cues across frequency [36], the evidence about a sound source location is accumulated by combining the log likelihood function across all N gammatone channels. In this way, the uncertainty associated with the azimuth estimate of a particular gammatone channel is taken into account, making an optimal use of the available information. This probabilistic integration of cues was also proposed by [21]. Gaussian mixtures are trained to recognize the binaural feature space in steps of 5°. To increase the localization accuracy, an exponential interpolation [29] is applied around the maximum of the log likelihood function accumulated across frequency in (11).

D. GMM Parameter Estimation

The most common approach to estimating the set of GMM parameters λ_{i,φ_k} is to use the iterative Expectation–Maximization (EM) algorithm [37]. After initializing the GMM parameters, each EM iteration consists of two steps, namely the E-step and the M-step, respectively. First, the E-step determines the membership of each training sample by assigning it to the Gaussian cluster which is most likely to have generated it. Based on the new membership estimation, the GMM parameters are recalculated in the M-step. This iterative procedure continues until the difference in likelihood between two successive iterations is less than a predefined threshold ϵ. Although Gaussian mixture models can, in theory, be established to approximate arbitrarily complex probability density functions, the quality and the robustness of the estimated GMM parameters depend on the number of Gaussian components and the way they are initialized prior to using the EM algorithm.

It is difficult to select the optimal number of Gaussian components K, because the "true" number is usually unknown. An extensive number of Gaussian components reduces the ability of the GMM model to generalize to observations which were not included in the training data. By choosing too few components, however, the essential characteristics of the feature space cannot be properly represented. One straightforward approach to choosing the number of GMM components is to visually identify the number of clusters, assuming that these clusters would also be recognized by the initialization procedure. Recently, an algorithm for unsupervised learning of Gaussian mixtures was presented, which automatically selects the optimal number of Gaussian components by minimizing a cost function based on the minimum description length (MDL) criterion [24]. Furthermore, the algorithm was reported to reduce the sensitivity of the EM algorithm to the initialization procedure by starting with significantly more components than required, and successively removing the unnecessary components using the MDL selection criterion. In this paper, both manual and automatic approaches were used to approximate the binaural feature space, of which the effect on localization performance will be compared in Section V.

The set of Gaussian parameters listed in (9) needs to be initialized prior to running the EM algorithm. A random initialization or the use of clustering algorithms is one of the common approaches [38]. In this paper, the k-means clustering algorithm [39] is used to find the initial parameters of the K Gaussian components. As the ITD and the ILD features have different scales, a variance normalization is performed prior to equalize both dimensions.

IV. Evaluation Setup

A. Baseline Systems

The proposed localization model is compared with three baseline systems by using the performance evaluation described in Section IV-D. All baseline algorithms were implemented to perform localization by using the same framing parameters as the proposed model: analysis window of 20 ms with a frame shift of 10 ms. The ITD estimates of all baseline systems were also refined by exponential interpolation [29].

GCC Gammatone: The first baseline system is based solely on the ITD analysis (1) with the same auditory front-end as the GMM-based localization model. Each local peak in the cross-correlation function is replaced by an impulse of the same height and convolved with a Gaussian kernel [6], [19]. The same parameters were used as suggested in [6]. This process sharpens the peaks in the cross-correlation function, and therefore is beneficial especially when there are multiple sources spatially close to one another. The final azimuth estimate is given by transforming the ITD according to a frequency-dependent mapping function, which takes into account the diffraction effects of the head and shoulders [6], [9], [19], and integrating the information across gammatone channels.

FFT PHAT and FFT SCOTM: The FFT-based generalized cross-correlation (GCC) technique is also used for comparison [12]. More specifically, two commonly used weighting functions for the GCC are explored: the phase transform (PHAT) [12] and a modified[2] version of the smoothed coherence transform (SCOTM) [40]. Let $\Phi_{ll}(\omega)$ and $\Phi_{rr}(\omega)$ be the auto power spectrum of a 20-ms time segment of the left and the right ear signal, respectively. Prior to spectral analysis, the time segments were multiplied by a Hamming window and padded with zeros to reach a window length corresponding to the next highest power of two. Furthermore, let $\Phi_{lr}(\omega)$ denote the cross power spectrum between the two signals. The frequency-specific weighting functions are given by $\Psi_{\text{PHAT}}(\omega) = 1/|\Phi_{rl}(\omega)|$ and $\Psi_{\text{SCOTM}}(\omega) = 1/[\Phi_{ll}(\omega)\Phi_{rr}(\omega)]^{0.3}$. Channel-dependent pre-whitening is applied to the binaural signal in the spectral domain prior to the GCC analysis, to reduce the dependence of localization on the structure of the source signal [23]. Similarly to the *GCC Gammatone*, a mapping function was employed to relate the broadband ITD estimate to the corresponding azimuth. The corresponding mapping functions were derived by learning the responses of the localization models to a speech source that was presented systematically at locations in the azimuth range of $[-50°, 50°]$.

B. GMM Settings

As discussed before, the GMMs were trained with the number of Gaussian components selected either manually [41] or automatically [24]. The automatic selection was constrained between $k_{\min} = 5$ and $k_{\max} = 25$ Gaussian components. In addition, the stopping criterion of the EM algorithm was set to $\epsilon = 1e^{-5}$ for both training methods with a maximum of 300 iteration steps.

C. Acoustic Conditions

Binaural cues were simulated at various locations in a room of dimensions $5.1 \times 7.1 \times 3$ m, as depicted in Fig. 2. The circles correspond to all possible receiver positions (KEMAR head) in the training phase. The diamonds represent the positions of the receiver at which the localization model was evaluated for various reverberation times. Note that only the receiver position 5 was used for both training and evaluation in order to study the influence of known/unknown receiver positions. The receiver was always oriented towards -90° and placed at 1.75 m above the ground, where the source azimuth was varied, at a radial distance of 1.5 m, with respect to receiver position. The positioning of sound sources (filled triangles) is sketched for one training and one evaluation scenario. The black triangle shows the placement of a target source at -50° with respect to receiver position 15, and the positions of an interfering source (crosses) placed at $\pm 5^\circ, \pm 10^\circ, \pm 20^\circ, \pm 30^\circ$ and $\pm 40^\circ$ relative to the target source, which were systematically processed in the multi-conditional training. As can be seen from Fig. 2, the maximum lateral sound source position had to be limited between $\pm 90^\circ$ since some receiver positions were too close to the room boundary (e.g., receiver position 11). On the other hand, the placement of the interfering source required an spatial offset of $\pm 40^\circ$ with respect to the target source position. Therefore, the GMM localization models were trained and evaluated for a narrower range of target azimuths between $\pm 50^\circ$ at every 5°, which resulted in 21 possible sound source locations. The gray triangles represent all 21 target positions with respect to receiver position 5 which were used for evaluation.

For evaluation, the surface *Acoustic plaster* was selected to characterize the reverberation of the room within the room simulation software [31], where the frequency-dependent absorption characteristic was applied for all room boundaries. In order to take into account mild-to-strong reverberation, several different sets of absorption coefficients were used for the room simulations and the frequency-dependent and the average reverberation time T_{60} for all experimental conditions are listed in Table I. Each row represents one experimental condition. Note that this reverberation characteristic is different from the one which was used to perform the multi-conditional training.

D. Performance Evaluation

The localization performance was evaluated on a frame-by-frame basis where an absolute error threshold $\theta_\varphi = 5^\circ$ was

[2] The square-root operator in the denominator of the conventional SCOT weighting was changed to cube-root compression. This modification was found to improve localization performance in reverberation.

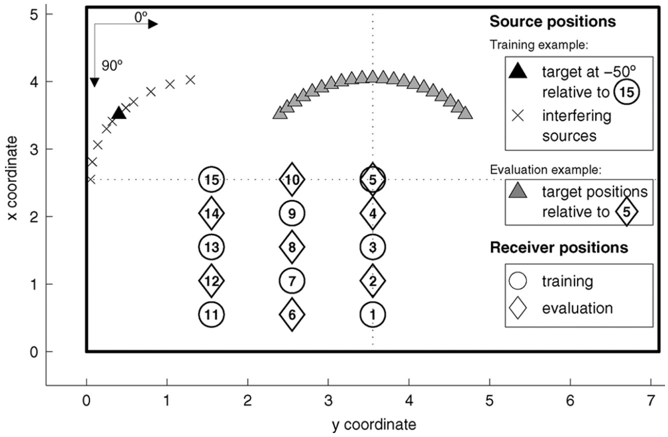

Fig. 2. Diagram showing the room dimension with all receiver positions used for training (circles) and for evaluation (diamonds). The triangles show exemplarily the positioning of sound sources for one training and one evaluation scenario. See text for details.

TABLE I
REVERBERATION TIMES T_{60} IN SECONDS FOR ALL EXPERIMENTAL CONDITIONS

	Frequency in Hz						
Surface	125	250	500	1000	2000	4000	mean
	0.48	0.33	0.13	0.08	0.06	0.06	0.19
Acoustic	0.67	0.48	0.24	0.15	0.11	0.10	0.29
plaster	0.81	0.61	0.36	0.23	0.18	0.15	0.39
	0.93	0.75	0.46	0.30	0.24	0.19	0.48
T_{60} in sec	1.09	0.89	0.56	0.39	0.30	0.24	0.58
	1.26	1.03	0.69	0.48	0.37	0.29	0.69

considered to classify the estimated source azimuth as either correct or anomalous [22], [42]. Correct estimates were further analyzed by means of the bias and the standard deviation. The performance of all the localization techniques was analyzed given the results of a series of Monte Carlo simulation experiments which were carried out for four sets of acoustic scenarios with 1, 2, 3, and 4 sources. Reference azimuth tracks were obtained from the anechoic speech files by using an energy-based VAD. Acoustic sources were represented by male speech selected from the TIMIT database [34], which were different from those used in the training stage. The average sentence length was 2.86 s. Regardless of the number of active sources in the acoustic mixture, each localization method produced one azimuth estimate per frame associated with the primary peak, while secondary peaks were ignored for a robust localization. In the case of a multisource mixture, the localization estimate was considered to be correct, only if the error was within the threshold θ_φ relative to one of the reference azimuth positions. Localization performance was evaluated at all 21 sound source positions within the azimuth range of $\pm 50^\circ$. The sound sources in a multisource mixture were positioned randomly, but the distance between nearby sources was constrained to at least 10°. The energy of each source was adjusted prior to spatialization to maintain a global SNR of 0 dB. All four scenarios, each consisting of 21 mixtures, were presented three times at all eight receiver positions selected for evaluation (see Fig. 2). Therefore, a total of 2016 ($4 \times 21 \times 3 \times 8$) acoustic mixtures were tested for each reverberation time.

TABLE II
EXPERIMENT 1: PERCENTAGE OF ANOMALOUS LOCALIZATION ESTIMATES DEPENDING ON THE NUMBER OF GAUSSIAN COMPONENTS

GMM	Reverberation time in seconds						
complexity	0.19	0.29	0.39	0.48	0.58	0.69	mean
Fixed 5	6.65	10.37	15.58	20.52	25.48	30.16	18.13
Fixed 11	2.20	4.30	8.05	12.28	17.16	22.03	11.00
Fixed 15	2.16	4.15	7.76	11.87	16.72	21.50	10.70
Fixed 21	2.11	4.06	7.71	11.77	16.62	21.40	10.61
Fixed 25	2.14	4.09	7.73	11.79	16.63	21.40	10.63
Fixed 31	2.12	4.09	7.76	11.85	16.67	21.52	10.67
Variable	2.14	4.05	7.70	11.76	16.61	21.39	10.61

V. LOCALIZATION EXPERIMENTS

A. *Experiment 1: Influence of GMM Model Complexity*

The first experiment investigated the influence of the GMM model complexity on localization performance. As described before, two different methods were used to determine the number of Gaussian components. First, the binaural feature space was approximated by a manually determined number of components, which was fixed across all azimuth angles and gammatone channels. The second method automatically selected the optimal model complexity for each azimuth angle and each gammatone channel independently, resulting in a variable model complexity.

The percentage of anomalies per frame is listed in Table II as a function of the GMM model complexity, where the performance was averaged across all four acoustic scenarios. With increasing number of Gaussian components, the model performed better in all reverberation conditions. In particular, the improvement was significant from 5 to 11 Gaussian components, but the performance gain started to saturate as the model complexity was further increased. For example, the average percentage of anomalies changed only by 0.09% between the order 15 and 21. In addition, the models with an extensive amount of Gaussian components (e.g., 25 or 31) performed slightly worse, which may indicate that the model was overtrained with the limited set of training data.

The last row in Table II shows the performance of the GMM model with variable model complexity. The average number of automatically determined Gaussian components was 17 across azimuth angle and gammatone channel and the performance was similar to the fixed GMM model with 21 components. However, the training procedure for this method takes significantly longer because the model is fitted for the whole complexity range between $k_{\min}$ and $k_{\max}$ Gaussian components. Furthermore, the similar performance of both procedures suggests that the requirement for learning the binaural feature space does not change across azimuth directions or gammatone channels. Thus, it seems to be no advantage in individualizing the training of the model for each azimuth direction and gammatone channel. Considering the performance saturation and the computational costs, the number of the GMM components was set to 15, constant across gammatone channels for the simulations presented in this study.

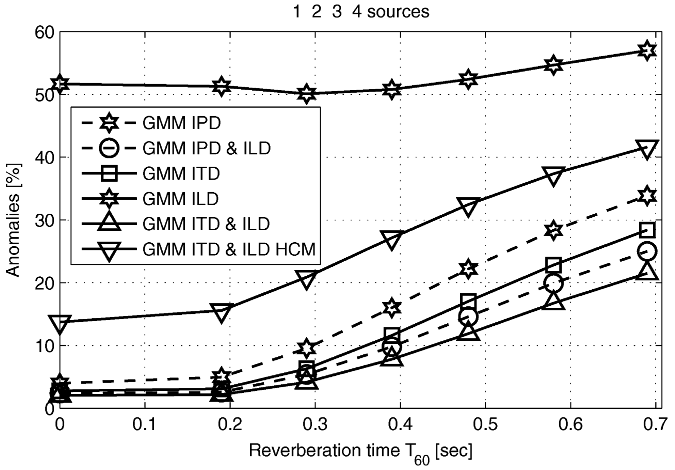

Fig. 3. Experiment 2: Effect of binaural cue selection on localization performance. Cues were extracted using either the gammatone-based front-end (solid lines) or the FFT-based representation (dashed lines). See text for details.

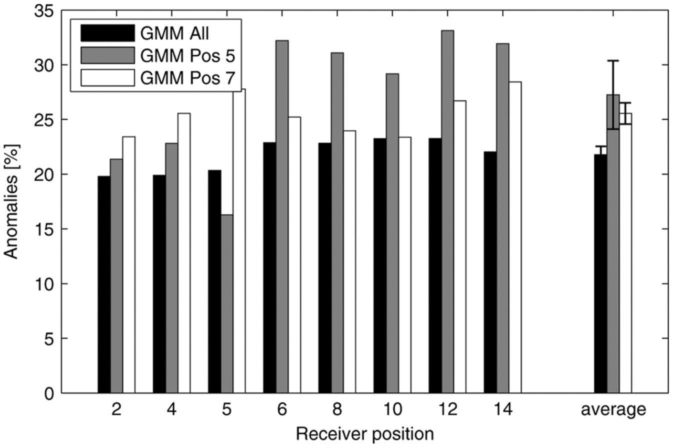

Fig. 4. Experiment 3: Percentage of anomalies evaluated at various receiver positions under reverberation condition ($T_{60} = 0.69$ s). The GMM localization model was trained using either one specific training position (Pos 5, Pos 7) or all eight training positions.

B. Experiment 2: Selection of Binaural Cues

The second experiment analyzed the impact of either using ITD or ILD only, or performing localization based on a joint two-dimensional binaural feature space. In Fig. 3, the percentage of anomalies averaged across four acoustic scenarios is shown as a function of the reverberation time. Also, the performance of the gammatone-based feature extraction (solid lines) defined in (6) is compared with a feature space obtained by an FFT-based auditory front-end (dashed lines), consisting of IPD and ILD, according to [21]. The same parameters were used to simulate the auditory periphery for the two implementations (see Section II-A), where no temporal smoothing was performed across frames.

The results show that the exclusive use of the interaural level cue (*GMM ILD*) is not sufficient to reliably determine the location of acoustic sources. Because the GMM models were trained for the reverberant environment, the localization performance slightly improved as the reverberation time increased up to $T_{60} = 0.3$ s, but overall, the error rate of *GMM ILD* was, in general above 50%. In contrast to *GMM ILD*, a reasonable localization performance could be achieved with the ITD model (*GMM ITD*). For example, even for a relatively long reverberation time of $T_{60} = 0.69$ s, the average percentage of anomalies was only about 28.39%.

It is apparent in Fig. 3 that the joint evaluation of both ITD and ILD produced the best result (*GMM ITD & ILD*). Although the isolated ILD cue does not allow for robust localization, it can significantly improve the localization performance by disambiguating the ITD information especially in acoustic conditions with strong reverberation.

Comparing the gammatone-based (solid lines) and the FFT-based auditory front-end (dashed lines), the GMM model using ITD performed noticeably better compared to the GMM using IPD only. The performance gap between the *GMM ITD* and the *GMM IPD* increased with the reverberation time, which may indicate that the azimuth-dependent IPD pattern is not so systematically modified by the reverberation time as the ITD pattern, and therefore, the GMM classifier cannot utilize it effectively. Using the joint feature space, the performance gap between the gammatone-based (*GMM ITD & ILD*) and FFT-based front-end (*GMM IPD & ILD*) was reduced, but the gammatone-based front-end performed consistently better than its counterpart.

So far, a basic neural transduction model for the inner hair cells was used for the simulation where the signals are half-wave rectified and square-root compressed. However, more detailed models typically employ a high-order low-pass filter to simulate the loss of phase-locking in the auditory nerve at higher frequencies [43], [44], as was used in this study for the model denoted *GMM ITD & ILD HCM*. Compared to the performance of the basic hair cell model (*GMM ITD & ILD*), that localization accuracy is significantly worse, which clearly reflects the importance of the ITD fine structure at higher frequencies for a better localization performance. Therefore, the basic hair cell model excluding the low-pass filter is used for the following experiments.

C. Experiment 3: Dependency on Source/Receiver Configuration

In the third experiment, the localization model was evaluated for the unknown source/receiver combinations. Two GMM models, denoted as *GMM Pos 5* and *GMM Pos 7*, were trained with binaural cues simulated at one specific receiver position (position 5 and 7, respectively), and compared to a model, *GMM All*, which was trained for all eight positions (1, 3, 5, 7, 9, 11, 13, and 15). The average percentage of anomalies for the receiver position considered for the evaluation is shown in Fig. 4, where the performance was averaged across all four acoustic scenarios (from one to four sources). To maximize the influence of receiver position on the binaural cues, long reverberation times (on average $T_{60} = 0.69$ s) were used in this experiment, but the performance was found to be similar regardless of the reverberation condition.

The model *GMM All* performed best among all three models, as shown in Fig. 4. Only at receiver position 5, the *GMM Pos 5* achieved a lower percentage of anomalies but the difference was rather small (4.19%) considering that the model had been trained for this very position. Indeed, the average performance was best with the *GMM All*, implying that the model is robust and can localize acoustic sources even from untrained receiver positions.

Overall, the *GMM Pos 5* produced the highest percentage of anomalies. The receiver position 5 is located in the center of the room, farthest from any room boundary. On the other hand, all the evaluation positions are relatively close to the room boundaries, where the pattern of the binaural cues can easily be modified or shifted by the acoustic reflection. For example, the performance is particularly low at receiver positions 6 and 12, which are close to the room boundary (see Fig. 2). Indeed, the model is quite sensitive to the placement of the receiver (the performance difference between position 5 and 12 is 17.58%), which obviously resulted in large variances shown by the error bars.

Compared to the receiver position 5 (see Fig. 2), position 7 is closer to the evaluation positions, and therefore the *GMM Pos 7* performed better than the *GMM Pos 5*. Especially at receiver position 6, 8, and 10, which are close to the model training position 7, the overall percentage of anomalies is almost as low as that for the *GMM All*. Nevertheless, the performance is, in general, better with the GMM model trained with multiple receiver positions than with one specific position.

So far, the radial distance between the acoustic sources and the receiver was kept constant at 1.5 m for both the training and the evaluation conditions. To analyze the effect of the radial distance on localization performance, the proposed localization model trained with binaural cues at a radial distance of 1.5 m was evaluated at receiver position 2 for five different radial distances.[3] The effect of distance is incorporated in the room simulation software [31] by simulating the distance-dependent circular wave attenuation and by modeling the air absorption with a low-pass filter.

The direct-to-reverberant ratio decreases with increasing radial distance, which may reduce the reliability of the measured binaural cues. The percentage of anomalies is presented in Fig. 5 averaged over all four acoustic scenarios. As shown in Fig. 5, the localization prediction of the model becomes less reliable with increasing source-receiver distance. With longer reverberation time, the localization error is more frequent, and the difference in performance is quite noticeable when going from 1.5 to 2.0 m. This result shows that the GMM model trained at a certain fixed distance may not successfully be applied for the localization prediction of more distant sources. However, the fact that the model performance improves at a shorter distance indicates that the model is capable to generalize to distances which have not been included in the training phase. The poorer performance at larger distances seems to be mainly determined by the reduced reliability of the binaural cues, resulting from the larger distance between the source and the receiver. This is in line with the expectation that the direct-to-reverberant ratio decreases with increasing radial distance and consequently affects the reliability of the binaural cues.

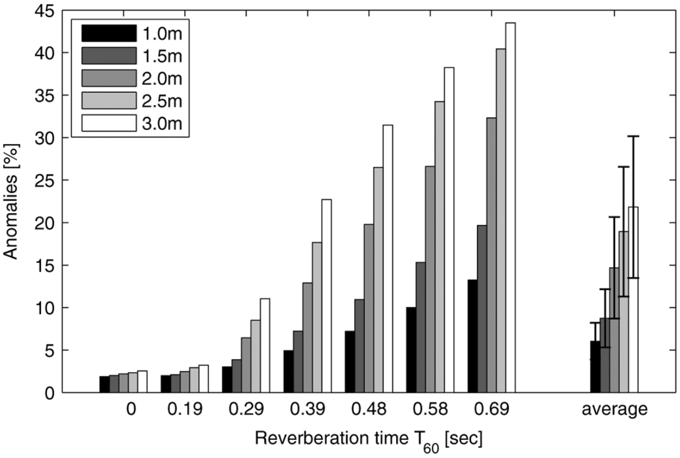

Fig. 5. Experiment 3: Percentage of anomalies depending on reverberation time evaluated at various distances between the source and receiver position 2. The GMM localization model was trained with binaural cues at a radial distance of 1.5 m using all eight training positions.

D. Experiment 4: Effect of the Number of Active Sources

Experiment 4 compared the performance of the GMM-based localization model to the baseline systems described in Section IV-A, where the effect of the number of active sources was also analyzed in terms of the localization performance. The localization results are presented in Fig. 6 for all evaluated systems as a function of the reverberation time T_{60}. While panels 6(a)–(d) show the results for all four acoustic scenarios, the percentage of anomalies and the standard deviation of the correct estimates are shown in panels 6(e) and (f), respectively.

As expected, the average percentage of anomalies increased with reverberation time for all localization methods. For the single source scenario shown in panel (a), the FFT-based model gave more accurate predictions than the gammatone-based GCC method, where the SCOTM weighting performed consistently better than the PHAT weighting. However, the performance of both FFT-based methods, SCOTM and PHAT, significantly deteriorated, when the number of sources increased.

The cross-correlation analysis based on the auditory front-end, *GCC Gammatone*, outperformed the FFT-based methods in all multisource conditions, which implies that the frequency selective cue extraction effectively resolved the dominant localization information of multiple sources. Nevertheless, the percentage of anomalies increased quite rapidly with the reverberation time.

The localization error was significantly reduced, when the ITD cue was employed in combination with Gaussian mixtures. Although the same information is available as in the case of *GCC Gammatone*, the presence of reverberation affected the performance less due to the multiconditional training and the probabilistic integration of source evidence across channels. In addition, localization performance was more robust in multisource scenarios because the binaural cues modified by competing sources were also considered in the multiconditional training phase.

As Fig. 6 clearly shows, the joint evaluation of ITD and ILD by Gaussian mixtures performed best in all acoustic scenarios, where the additional ILD information was especially important

[3] Whereas ITDs can be considered to be fairly independent of the distance between the source and the receiver, ILDs can change considerably with distance in the proximal region, in particular for nearby sources at distances below 0.5 m [45]. However, for distances larger then 1 m, the distance-dependent changes of binaural cues are assumed to be negligible [46].

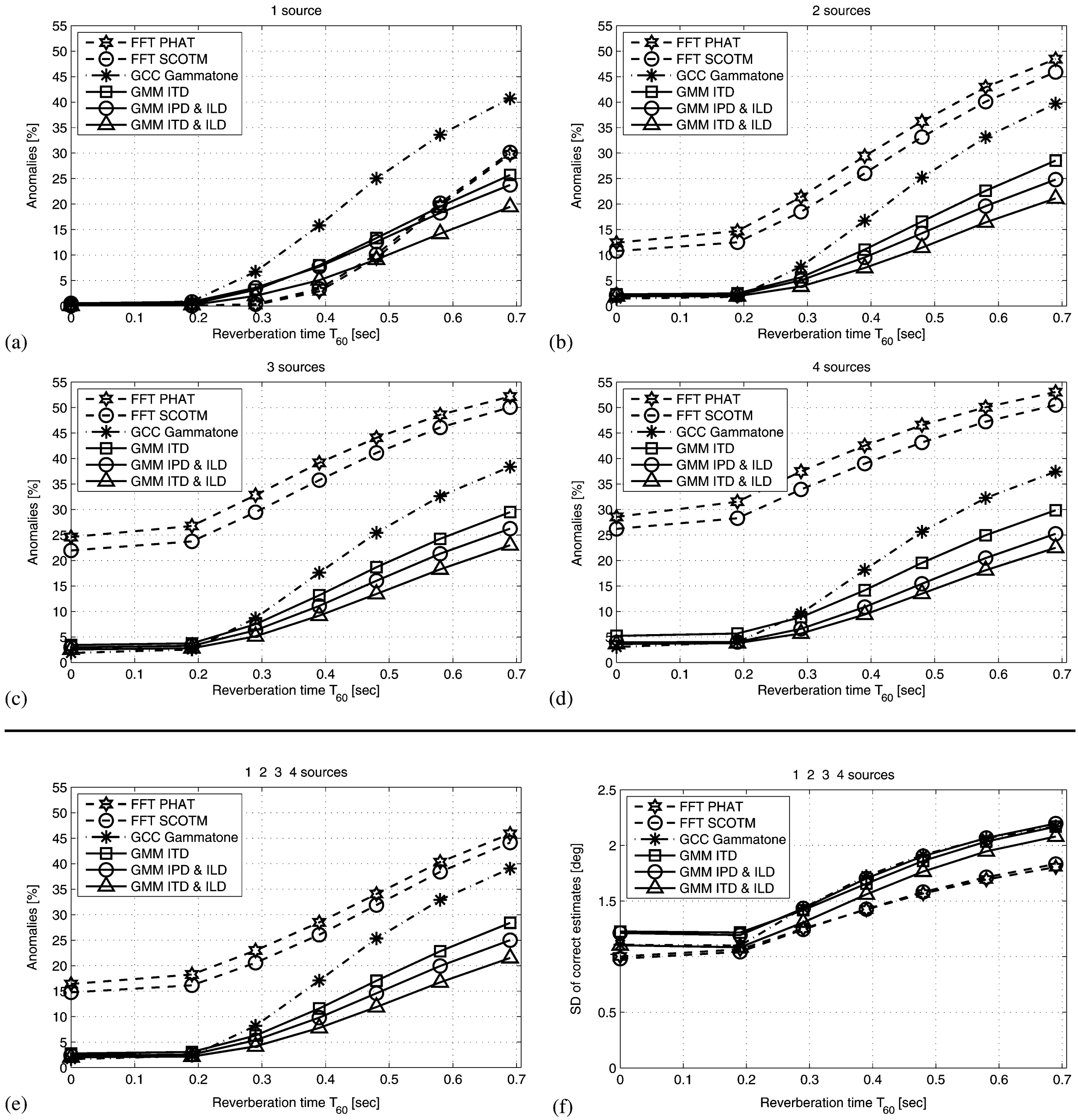

Fig. 6. Experiment 4: Percentage of anomalies depending on reverberation time T_{60} of all baseline methods and GMM-based localization algorithms evaluated in four acoustic scenarios (a)–(d), consisting of 1, 2, 3, and 4 sources. Moreover, the percentage of anomalies and the standard deviation of the correct estimates summarized across all four acoustic scenarios are given in (e) and (f), respectively. Results are shown for all three categories of localization methods, namely the FFT-based methods (dashed line), the gammatone-based GCC (dash-dotted lines) and the GMM-based models (solid lines).

to cope with strong reverberation. Indeed, the performance of the FFT-based methods was strongly affected by the number of active sources, for which the GMM model using both ITD and ILD cues was almost independent.

E. Experiment 5: Blind Estimation of the Number of Acoustic Sources

The number of active sources in an acoustic mixture can be a valuable information, which might be used for blind source separation algorithms, to control and steer the beams of a microphone array, or to post-process the frame-by-frame localization estimates. In Experiment 5, the capability of the proposed localization method to predict the number of active sources in an acoustic mixture was explored. Therefore, a histogram based on the frame-by-frame azimuth estimates is computed with a resolution of 5° by pooling the azimuth estimates of all frames over the entire acoustic mixture. After normalization, all local peaks in the histogram which fall below a predefined threshold θ_h are discarded. The remaining histogram peaks are assumed to be caused by active sources and are therefore selected as source candidates. The following performance measure is used to take into account errors which are caused by either selecting more or fewer source candidates than the true number of active sources. Let $r_\varphi = \{r_{\varphi_1}, \ldots, r_{\varphi_R}\}$ be a set of R reference source po-

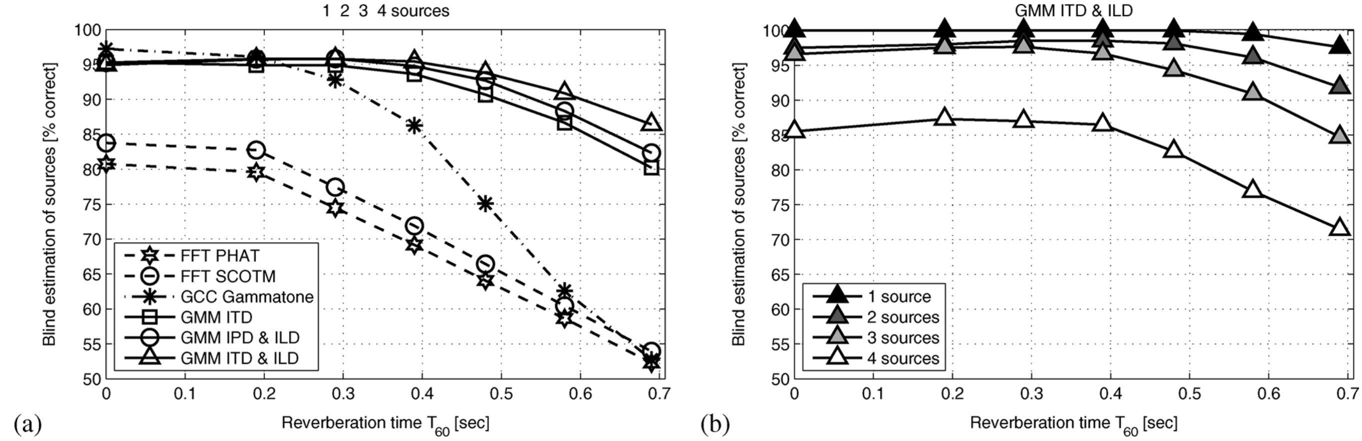

Fig. 7. Experiment 5: Performance of estimating the number of active sources in mixtures consisting of 1, 2, 3, and 4 sources as a function of reverberation time T_{60}. In (a), the average performance is presented for all baseline methods and GMM-based localization algorithms. In (b), the dependence of performance on the number of sources is shown for the GMM-based evaluation of both ITD and ILD.

sitions which were present simultaneously in the acoustic mixture, and let $\hat{r}_\varphi = \{\hat{r}_{\varphi_1}, \ldots, \hat{r}_{\varphi_{\hat{R}}}\}$ be a set of $\hat{R}$ estimated source candidates. The percentage of correctly identified number of sources is given by

$$p_{\mathrm{c}} = 100 \cdot \frac{|r_\varphi \cap \hat{r}_\varphi|}{|r_\varphi \cup \hat{r}_\varphi|}. \tag{12}$$

The intersection of the reference source positions r_φ and the estimated source candidates $\hat{r}_\varphi$ is related to the union of the two. The operator $|\cdot|$ represents the cardinality of a set, which is a measure of the number of elements of a set. In this way, performance decreases if more than the true number of sources are selected, although all reference positions might have been correctly identified.

Performance was evaluated for various histogram thresholds θ_h, and results are reported for the threshold which gave the overall best result for each localization method independently. The percentage of correctly identified number of sources is shown in Fig. 7(a). Performance was averaged over four acoustic scenarios, containing between one and four sources and over all eight evaluation positions. As already shown in experiment 4, the presence of multiple sources has a severe effect on the performance of the FFT-based localization methods. Even under anechoic conditions, the performance of estimating the number of active sources was below 85%. With increasing reverberation time, performance decreased to about 55%. Again, the SCOTM filtering performed consistently better than the PHAT weighting. Up to a reverberation time of $T_{60} = 0.2$ s, the gammatone-based GCC method is slightly superior to the GMM-based models, but with stronger reverberation, the performance of the GCC-based method rapidly decreased to 53% at a reverberation time of $T_{60} = 0.69$ s. Due to the multiconditional training of the Gaussian mixtures, localization performance is robust against the impact of reverberation. As a consequence, the GMM based models led to a cleaner histogram of source positions, which allowed for a more reliable identification of the number of active sources. Even at a reverberation time of $T_{60} = 0.69$ s, about 86% of the multisource mixtures were correctly classified using the *GMM ITD & ILD*. In Fig. 7(b), the performance of *GMM ITD & ILD* is presented depending on the number of acoustic sources. Whereas classification was quite robust for 1,2, and 3 sources, performance significantly decreased for mixtures consisting of four sources.

VI. Discussion and Conclusion

A robust acoustic localization model was presented, which is based on the supervised learning of azimuth-dependent binaural feature maps consisting of ITD and ILD. The model was evaluated in simulated adverse acoustic scenarios and outperformed state-of-the-art localization techniques. Furthermore, the model was capable of generalizing to unknown source/receiver configurations which were not included in the training stage. Based on the frame-by-frame localization estimates, an efficient histogram technique allowed to robustly estimate the number of active sources in acoustic multi-source mixtures.

The robustness of the model against reverberation and the presence of multiple sources is attributed to three factors. First, due to the auditory front-end, the frequency-dependent, dominant localization information of multiple sources can be spectrally resolved, allowing for a robust estimation of the binaural cues. Second, GMMs are used to evaluate the joint binaural feature space and to accumulate evidence of possible source locations across frequency in a probabilistic way, taking into account the available information in an optimum way. Third, the multiconditional training incorporates the uncertainty of the binaural cues caused by room effects, reverberation, the presence of multiple sources, and changes in the source/receiver configurations.

It was shown that integrating the probabilities of possible sound source locations across frequency is superior to accumulating the localization cues directly. Moreover, the joint evaluation of ITD and ILD disambiguates the information derived from the ITD cue, especially in strong reverberation, increasing the robustness of the localization model.

Although the concept of using either IPD or ITD might provide similar information, the localization performance using the ITD cue was superior to the IPD cue. This comparison is based on the assumption that the distribution of both ITD and IPD can be modeled equally well using a sum of Gaussian distributions. One possible explanation might be that the multiple clusters in the ITD feature space, which reflect the ambiguous ITD information, are warped to the interval between $\pm\pi$ in the IPD representation. Since the centered ITD clusters are more sharp and the more lateral clusters are more broad, this differentiated analysis is lost in the IPD representation, presumably decreasing the localization performance. Furthermore, it is worthwhile to note that the GMM localization model using a basic hair cell model outperformed the more detailed hair cell model including higher order low-pass filtering. Thus, the probabilistic model is capable of exploiting the ITD in the fine structure at higher frequencies, which is generally accepted not to be accessible by the human auditory system [47], [48]. Thus, the model is not strictly limited by the processing which is believed to be performed by the human auditory system.

The broadband FFT localization (PHAT and SCOTM) is prone to errors in multisource scenarios, because it actually averages the directional cues of all sources over a short time segment, which can lead to a phantom source that does not necessarily reflect the source position of the most dominant source. This is especially likely to happen if the energetic contribution of sources is equally strong and the sources are located symmetrically but in opposite directions with respect to the receiver (e.g., -40° and 40°). Furthermore, a noticeably longer analysis window than 20 ms is commonly used to increase the robustness in adverse conditions [23]. However, this is only reasonable for single-source localization.

The current analysis of the localization model was restricted to the frontal horizontal plane, whereas front-back discrimination and localization in the elevation domain are tasks for further investigations. Using the framework of Gaussian mixtures, the binaural feature space could be readily extended by additional descriptive features which are depending on the position of sound sources. For example, in order to extend the model to the elevation domain, the use of spectral cues might be beneficial [49].

The radial distance between the source and the receiver, which determines the relation between the direct and the reverberated sound, was a sensitive parameter. Similar to the reverberation time, the radial distance is a source of uncertainties which modifies the distribution of the binaural cues. Localization performance significantly decreased at larger radial distances, which is in line with behavioral data observed for humans [50]. In order to improve the working range of the model, it might be beneficial to train the model either with binaural cues simulated at a larger radial distance or with binaural cues corresponding to various radial distances.

The reported localization performance was achieved by applying the probabilistic model on a frame-by-frame basis, accumulating evidence over frequency channels. However, accumulating evidence of sound source locations across a larger time span could further increase localization performance. Integrating the localization cues over patches across time and frequency, which are believed to belong to a single source was reported to significantly improve ITD-based localization performance [51]–[53]. Instead of integrating the localization cues directly, the proposed probabilistic localization model could combine likelihoods of sound source locations across patches.

Due to its robustness and high temporal resolution, the localization model presented in this study might be very suitable as a front-end for CASA algorithms that segregate and recognize sound sources in complex acoustic mixtures.

Acknowledgment

The authors would like to thank Dr. A. Härmä, Dr. M. H. Park, and the anonymous reviewers for their useful comments that have greatly improved the clarity of this paper. They would also like to thank the Perception and Neurodynamics Laboratory (PNL), The Ohio State University, Columbus, for making their source code freely available.

References

[1] J. DiBiase, H. Silverman, and M. Brandstein, , M. Brandstein and D. Ward, Eds., "Robust localization in reverberant rooms," in *Microphone Arrays: Signal Processing Techniques and Applications*. Berlin, Germany: Springer-Verlag, 2001, ch. 8, pp. 157–180.

[2] M. L. Hawley, R. Y. Litovsky, and H. S. Colburn, "Speech intelligibility and localization in a multi-source environment," *J. Acoust. Soc. Amer.*, vol. 105, no. 6, pp. 3436–3448, Jun. 1999.

[3] A. W. Bronkhorst and R. Plomp, "Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing," *J. Acoust. Soc. Amer.*, vol. 92, no. 6, pp. 3132–3139, Dec. 1992.

[4] A. S. Bregman*, Auditory Scene Analysis.*. Cambridge, MA: MIT Press, 1990.

[5] *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, D. L. Wang and G. Brown, Eds. New York: IEEE Press/ Wiley-Interscience, 2007.

[6] 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.*, vol. 43, no. 4, pp. 361–378, 2004.

[7] 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.*, vol. 14, no. 1, pp. 58–67, Jan. 2006.

[8] A. Härmä, J. Jakka, M. Tikander, M. Karjalainen, T. Lokki, J. Hiipakka, and G. Lorho, "Augmented reality audio for mobile and wearable appliances," *J. Audio Eng. Soc.*, vol. 52, no. 6, pp. 618–639, Jun. 2004.

[9] M. Bodden, "Modeling human sound-source localization and the cocktail-party-effect," *Acta Acust.*, vol. 1, no. 1, pp. 43–55, 1993.

[10] C. Faller and J. Merimaa, "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *J. Acoust. Soc. Amer.*, vol. 116, no. 5, pp. 3075–3089, Nov. 2004.

[11] K. Wilson and T. Darrell, "Improving audio source localization by learning the precedence effect," in *Proc. ICASSP*, 2005, vol. 4, pp. 18–23.

[12] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, no. 4, pp. 320–327, Aug. 1976.

[13] L. A. Jeffress, "A place theory of sound localization," *J. Comput. Physiol. Psychol.*, vol. 41, no. 1, pp. 35–39, Feb. 1948.

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

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

[16] J. Braasch, "Modelling of binaural hearing," in *Communication Acoustics*, J. Blauert, Ed. New York: Springer, 2005, ch. 4, pp. 75–108.

[17] H. Viste and G. Evangelista, "Binaural source localization," in *Proc. DAFx-04*, 2004, pp. 145–150.

[18] B. G. Shinn-Cunningham, N. Kopco, and T. J. Martin, "Localizing nearby sound sources in a classroom: Binaural room impulse responses," *J. Acoust. Soc. Amer.*, vol. 117, no. 5, pp. 3100–3115, May 2005.

[19] N. Roman, D. L. Wang, and G. J. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Amer.*, vol. 114, no. 4, pp. 2236–2252, Oct. 2003.

[20] G. Brown, S. Harding, and J. Barker, "Speech separation based on the statistics of binaural auditory features," in *Proc. ICASSP*, 2006, vol. 5, pp. 14–19.

[21] J. Nix and V. Hohmann, "Sound source localization in real sound fields based on empirical statistics of interaural parameters," *J. Acoust. Soc. Amer.*, vol. 119, no. 1, pp. 463–479, Jan. 2006.

[22] B. Champagne, S. Bedard, and A. Stephenne, "Performance of time-delay estimation in the presence of room reverberation," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 2, pp. 148–152, Mar. 1996.

[23] 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*, vol. 1, pp. 25–36, 2005.

[24] M. Figueiredo and A. Jain, "Unsupervised learning of finite mixture models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 3, pp. 381–396, Mar. 2002.

[25] J. Blauert, *Spatial Hearing—The Psychophysics of Human Sound Localization*. Cambride, MA: MIT Press, 1997.

[26] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Comput. Speech Lang.*, vol. 8, pp. 297–336, 1994.

[27] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.*, vol. 47, no. 1–2, pp. 103–138, Aug. 1990.

[28] B. C. J. Moore, B. R. Glasberg, and T. Baer, "A model for prediction of thresholds, loudness, and partial loudness," *J. Audio Eng. Soc.*, vol. 45, no. 4, pp. 224–240, Apr. 1997.

[29] L. Zhang and X. Wu, "On cross correlation based discrete time delay estimation," in *Proc. ICASSP*, 2005, vol. 4, pp. 981–984.

[30] G. Jacovitti and G. Scarano, "Discrete time techniques for time delay estimation," *IEEE Trans. Signal Process*, vol. 41, no. 2, pp. 525–533, Feb. 1993.

[31] D. R. Campbell, K. J. Palomäki, and G. Brown, "A MATLAB simulation of "shoebox" room acoustics for use in research and teaching," *Comput. Inf. Syst.*, vol. 9, no. 3, pp. 48–51, 2005.

[32] W. Gardner and K. Martin, "HRTF measurements of a KEMAR dummy-head microphone," MIT Media Lab Perceptual Comput. Tech. Rep. #280, 1994.

[33] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[34] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," National Inst. of Standards and Technol., Gaithersburg, MD, Tech. Rep. NISTIR 4930, 1993.

[35] D. Reynolds and R. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 72–83, Jan. 1995.

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

[37] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood estimation from incomplete data via the EM algorithm," *J. R. Statist. Soc. B*, vol. 39, no. 1, pp. 1–38, 1977.

[38] G. McLachlan and D. Peel, *Finite Mixture Models*. New York: Wiley, 2000.

[39] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982.

[40] G. C. Carter, A. H. Nuttall, and P. G. Cable, "The smoothed coherence transform," *Proc. IEEE*, vol. 61, no. 10, pp. 1497–1498, Oct. 1973.

[41] I. T. Nabney and C. M. Bishop, "NETLAB package," Aug. 9, 2008. [Online]. Available: http://www.ncrg.aston.ac.uk/netlab/index.php

[42] J. P. Ianniello, "Time delay estimation via cross-correlation in the presence of large estimation errors," *IEEE Trans. Acoust., Speech, Signal Process*, vol. 30, no. 6, pp. 998–1003, Dec. 1982.

[43] L. R. Bernstein and C. Trahiotis, "The normalized correlation: Accounting for binaural detection across center frequency," *J. Acoust. Soc. Amer.*, vol. 100, no. 6, pp. 3774–3784, Dec. 1996.

[44] L. R. Bernstein, S. van de Par, and C. Trahiotis, "The normalized interaural correlation: Accounting for NoSπ thresholds obtained with Gaussian and "low-noise" masking noise," *J. Acoust. Soc. Amer.*, vol. 106, no. 2, pp. 870–876, Aug. 1999.

[45] D. S. Brungart and W. M. Rabinowitz, "Auditory localization of nearby sources. head-related transfer functions," *J. Acoust. Soc. Amer.*, vol. 106, no. 3, pp. 1465–1479, Sep. 1999.

[46] D. S. Brungart, N. I. Durlach, and W. M. Rabinowitz, "Auditory localization of nearby sources. II: Localization of a broadband source," *J. Acoust. Soc. Amer.*, vol. 106, no. 4, pp. 1956–1968, Oct. 1999.

[47] R. Klumpp and H. Eady, "Some measurements of interaural time difference thresholds," *J. Acoust. Soc. Amer.*, vol. 28, no. 5, pp. 859–860, Sep. 1956.

[48] J. Zwislocki and R. Feldman, "Just noticeable differences in dichotic phase," *J. Acoust. Soc. Amer.*, vol. 28, no. 5, pp. 860–864, Sep. 1956.

[49] P. Zakarauskas and M. S. Cynader, "A computational theory of spectral cue localization," *J. Acoust. Soc. Amer.*, vol. 94, no. 3, pp. 1323–1331, Sep. 1993.

[50] S. Devore, A. Ihlefeld, B. G. Shinn-Cunningham, and B. Delgutte, "Neural and behavioral sensitivities to azimuth degrade with distance in reverberant environments," in *Hearing - From Sensory Processing to Perception*, B. Kollmeier, G. Klump, V. Hohmann, U. Langemann, M. Mauermann, S. Uppenkamp, and J. Verhey, Eds. New York: Springer, 2007, pp. 505–516.

[51] H. Christensen, N. Ma, S. N. Wrigley, and J. Barker, "Integrating pitch and localisation cues at a speech fragment level," in *Proc. Interspeech*, 2007, pp. 2769–2772.

[52] H. Christensen, N. Ma, S. N. Wrigley, and J. Barker, "Improving source localisation in multi-source, reverberant conditions: Exploiting local spectro-temporal location cues," *J. Acoust. Soc. Amer.*, vol. 123, no. 5, pp. 3294(A)–, Jul. 2008.

[53] H. Christensen, N. Ma, S. N. Wrigley, and J. Barker, "A speech fragment approach to localising multiple speakers in reverberant environments," in *Proc. ICASSP*, 2009, pp. 4593–4596.

Tobias May received the Dipl.-Ing. (FH) degree in hearing technology and audiology from the Oldenburg University of Applied Science, Oldenburg, Germany, in 2005 and the M.Sc. degree in hearing technology and audiology from the University of Oldenburg, Oldenburg, Germany, in 2007. He is currently pursuing the Ph.D. degree. Since 2007, he has been with the Eindhoven University of Technology, The Netherlands. Since 2010, he is affiliated with the University of Oldenburg.

His research interests include computational auditory scene analysis, binaural signal processing, and automatic speaker recognition.

Steven van de Par studied physics at the Eindhoven University of Technology, Eindhoven, The Netherlands, and received the Ph.D. degree from the Eindhoven University of Technology, in 1998, on a topic related to binaural hearing.

As a Postdoctoral Researcher at the Eindhoven University of Technology, he studied auditory–visual interaction and was a Guest Researcher at the University of Connecticut Health Center. In early 2000, he joined Philips Research, Eindhoven, to do applied research in digital signal processing and acoustics. His main fields of expertise are auditory and multisensory perception, low-bit-rate audio coding and music information retrieval. He has published various papers on binaural auditory perception, auditory–visual synchrony perception, audio coding, and music information retrieval (MIR)-related topics. Since April 2010, he has held a professor position in acoustics at the University of Oldenburg, Oldenburg, Germany.

Armin Kohlrausch studied physics at the University of Göttingen, Göttingen, Germany, specializing in acoustics and received the M.S. degree in 1980 and the Ph.D. degree in 1984, both in perceptual aspects of sound from the University of Göttingen.

From 1985 until 1990, he worked at the Third Physical Institute, University of Göttingen, and was responsible for research and teaching in the fields psychoacoustics and room acoustics. In 1991, he joined Philips Research Laboratories, Eindhoven, and worked in the Speech and Hearing Group of the Institute for Perception Research (IPO). Since 1998, he has combined his work at Philips Research Laboratories with a Professor position for multisensory perception at the TU/e. In 2004, he was appointed a Research Fellow of Philips Research. He is a member of a great number of scientific societies, both in Europe and the U.S.

Dr. Kohlrausch has been a Fellow of the Acoustical Society of America since 1998 covering the areas of binaural and spatial hearing. His main scientific interest is in the experimental study and modeling of auditory and multisensory perception in humans and the transfer of this knowledge to industrial media applications.