Binaural Localization of Multiple Sources in Reverberant and Noisy Environments

John Woodruff, *Student Member, IEEE*, and DeLiang Wang, *Fellow, IEEE*

***Abstract*—Sound source localization from a binaural input is a challenging problem, particularly when multiple sources are active simultaneously and reverberation or background noise are present. In this work, we investigate a multi-source localization framework in which monaural source segregation is used as a mechanism to increase the robustness of azimuth estimates from a binaural input. We demonstrate performance improvement relative to binaural only methods assuming a known number of spatially stationary sources. We also propose a flexible azimuth-dependent model of binaural features that independently captures characteristics of the binaural setup and environmental conditions, allowing for adaptation to new environments or calibration to an unseen binaural setup. Results with both simulated and recorded impulse responses show that robust performance can be achieved with limited prior training.**

***Index Terms*—Binaural sound localization, computational auditory scene analysis (CASA), monaural grouping, reverberation.**

I. Introduction

LOCALIZATION of multiple sound sources from a binaural input is a challenging problem that has applications in hearing prostheses, spatial sound reproduction, and mobile robotics. Binaural localization has received significant attention in the field of computational auditory scene analysis (CASA) [37], which is guided by principles in the perceptual organization of sound by human listeners. Two principal localization cues are interaural time difference (ITD), also commonly referred to as the time difference of arrival, and interaural level difference (ILD), which is due to the effects of the head, torso, and outer ear [5].

Many of the key differences between localization methods result from assumptions about environmental factors such as source propagation, background noise, or the microphone setup. The generalized cross correlation (GCC) method is a well-known approach for ITD estimation that assumes ideal single-path propagation [22]. As this model does not account for the effect of reverberation or background noise, techniques have been proposed to make GCC more robust in reverberation [7], [35] or to more accurately model source propagation in reverberant [3] and noisy [13] environments. While these approaches are applicable to any two-microphone setup, there has been substantial research on localization models that are tailored to a binaural setup [36]. Recent efforts have incorporated azimuth-dependent models of ITD and ILD [27], [32], [38], and it was shown in [27] that jointly considering ITD and ILD improves azimuth estimation relative to ITD alone. However, because the frequency-dependent pattern of ITDs and ILDs can vary considerably across individuals, azimuth-dependent models require prior training or calibration with the binaural input and may suffer performance degradation for different binaural setups.

Methods also differ in how interaural information is integrated across time and frequency. These differences are largely a function of assumptions about source activity and interaction. In the case of multiple moving sources, statistical tracking approaches have been proposed to propagate localization estimates across time [12], [25], [33]; however, binaural methods have focused on conditions with little reverberation or background noise [12], [33]. If it can be assumed that sources are spatially stationary over a given interval of time, a simple approach is to first integrate azimuth information across frequency, then average across time and select multiple peaks in the resulting azimuth-dependent response function [24], [34]. Methods based on histograms of frame-level azimuth estimates have also been proposed [1], [27]. These methods assume that each active source will be dominant in a sufficient number of frames. This approach can be effective if there is sufficient azimuth separation and time integration, but can perform poorly when one source is dominant over the majority of the integration period. By assuming source sparsity in a time–frequency (T-F) representation, spatial clustering methods have been proposed to jointly segregate and localize a known number of spatially stationary sources [26], [30]. In this case, localization could potentially be improved by integrating features over a subset of T-F units, however the demonstrated benefit of recent systems is in terms of segregation rather than localization [26].

In this work, we propose a localization method where, similar to spatial clustering methods, azimuth estimates are derived from only those T-F units in which a given source is thought to be dominant. In contrast to existing spatial clustering methods, segregation is performed on the basis of both monaural and binaural cues and we demonstrate that this improves azimuth estimation in reverberant and noisy conditions. The proposed approach is motivated by psychoacoustics studies showing that grouping on the basis of monaural cues can influence localization judgements by human listeners (see, e.g., [4]), and by prior

Manuscript received March 29, 2011; revised October 25, 2011; accepted December 31, 2011. Date of publication January 11, 2012; date of current version March 14, 2012. This work was supported in part by the Air Force Office of Scientific Research (AFOSR) under Grant FA9550-08-1-0155, in part by the National Science Foundation (NSF) under Grant IIS-0534707, and a grant from the Oticon Foundation. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jingdong Chen.

J. Woodruff is with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH, 43210-1277 USA (e-mail: woodrufj@cse.ohio-state.edu).

D. L. Wang is with the Department of Computer Science and Engineering and the Center for Cognitive Science, The Ohio State University, Columbus, OH, 43210-1277 USA (e-mail: dwang@cse.ohio-state.edu).

Digital Object Identifier 10.1109/TASL.2012.2183869

work in which pitch-based grouping was shown to improve localization and segregation of voiced speech [39]. Here we extend our previous system [39] to deal with both voiced and unvoiced speech and develop a novel azimuth-dependent binaural model of interaural cues. We provide extensive evaluation of the localization procedure in adverse conditions and in situations with limited prior knowledge of the binaural setup.

Some existing work has incorporated monaural features within a two-microphone localization or tracking framework. The methods in [21], [28] perform joint estimation of azimuth and pitch, while [29] considers joint tracking of azimuth and monaural spectral properties. These studies, however, have focused either on the single source case, or on conditions without noise or reverberation. In [7] and [10], pitch information is used to improve frame-level ITD estimation of a dominant source in reverberation. Under the assumption that sources have strong harmonic components, a method to cluster ITD cues extracted from sinusoidal tracks is proposed in [40].

In Section II, we describe extraction of binaural features and propose a novel azimuth-dependent binaural model and associated training procedure. We summarize the monaural CASA methods used in Section III. In Section IV, we describe how binaural and monaural cues are integrated within the proposed framework for the purpose of multi-source localization. We describe the evaluation methodology in Section V and discuss the results of several experiments using both simulated and measured binaural impulse responses in Section VI. Section VII concludes the paper with a discussion of the insights gained from the evaluation and future work.

II. Binaural Pathway

A. Auditory Periphery and Binaural Feature Extraction

We assume a binaural input signal sampled at a rate of 44.1 kHz. The binaural signal is analyzed using a bank of 64 gammatone filters [31] with center frequencies from 80 to 5000 Hz spaced on the equivalent rectangular bandwidth scale. Each bandpass filtered signal is divided into 20-ms time frames with a frame shift of 10 ms to create a cochleagram [37] of T-F units. A T-F unit is an elemental sound component from one frame, indexed by m, and one filter channel, indexed by c. We denote a T-F unit as $u^E_{c,m}$ where $E \in \{L, R\}$ indicates the left or right ear signal.

The binaural pathway consists of a low-level feature extraction stage where we calculate the ITD, denoted $\tau_{c,m}$, and ILD, denoted $\lambda_{c,m}$, for each T-F unit pair. We calculate ITD as the maximum peak in a running cross-correlation between T-F units $u^L_{c,m}$ and $u^R_{c,m}$, where we consider time lags between -1 and 1 ms. ILD corresponds to the energy ratio in dB between $u^L_{c,m}$ and $u^R_{c,m}$. Both values are calculated as described in [39]. We then map ITD-ILD value pairs to azimuth-dependent features using the trained probabilistic models described below.

B. Azimuth-Dependent Binaural Model

We employ a simple and flexible azimuth-dependent Gaussian mixture model (GMM) of ITD and ILD. The model independently captures the frequency-dependent pattern of ITD and ILD values due to direct-path propagation, which we refer to as *direct-path cues*, and the statistical effect of environmental factors such as noise and reverberation. As a result, the model is easily adaptable to different binaural setups and acoustic conditions.

We denote the azimuth-dependent model of ITD and ILD as $P_c(\tau, \lambda|\theta)$, which represents the likelihood of observing a pair of ITD and ILD values in frequency channel c given energy from a point source with azimuth θ. In order to model the direct-path ITD and ILD independently of variance due to the acoustic conditions, we introduce the direct-to-residual ratio (DRR) for a point source as a latent variable. We calculate DRR, denoted $r_{c,m}$, within a pair of T-F units $u^L_{c,m}$ and $u^R_{c,m}$ as

$$r_{c,m} = \frac{\sum_n \left(x^L_{c,m}[n]^2 + x^R_{c,m}[n]^2\right)}{\sum_n \left(x^L_{c,m}[n]^2 + x^R_{c,m}[n]^2 + v^L_{c,m}[n]^2 + v^R_{c,m}[n]^2\right)} \tag{1}$$

where n indexes a signal sample, $x^E_{c,m}$ denotes the component of $u^E_{c,m}$ in response to the direct-path of the target source, and $v^E_{c,m} = u^E_{c,m} - x^E_{c,m}$. Each summation is over the interval of the corresponding T-F unit. Note that our use of DRR differs from the common use as an acronym for direct-to-reverberant ratio.

Given the DRR, r, and the direct-path ITD and ILD associated with azimuth θ, denoted τ_θ and λ_θ, we approximate the joint ITD-ILD observation likelihood for an individual frequency channel using

$$P_c(\tau, \lambda|\theta) \approx \sum_r P_c(\tau|r, \tau_\theta) P_c(\lambda|r, \lambda_\theta) P_c(r) \tag{2}$$

where $P_c(r)$ denotes the prior probability of DRR. Here, we assume that r is independent of τ_θ and λ_θ and that the observed ITD and ILD values are conditionally independent given the DRR and direct-path cues. We also approximate integration over r by summation over a discrete set of values.

Due to spatial aliasing, the probability space for observed ITDs in higher frequency channels is multi-modal. We therefore use a mixture of Gaussians to capture $P_c(\tau|r, \tau_\theta)$, or

$$P_c(\tau|r, \tau_\theta) = \sum_{k=1}^{K_c} \rho_{c,k}(r, \tau_\theta) \mathcal{N}(\tau|\mu_{c,k}(r, \tau_\theta), \sigma_{c,k}(r, \tau_\theta)) \tag{3}$$

where K_c is determined based on the channel center frequency, the direct-path ITD, and the range of observable ITD values (between -1 and 1 ms in this study). The ILD likelihood is well described by a single Gaussian, $P_c(\lambda|r, \lambda_\theta) = \mathcal{N}(\lambda|\mu_c(r, \lambda_\theta), \sigma_c(r, \lambda_\theta))$. Finally, letting R be the number of discretized values for r, $P_c(r)$ is a set of R scalar values. Given that each component of the model is either a set of Gaussians or a scalar, the full model can be written as a two-dimensional GMM with $R \cdot K_c$ components.

We show example models for $\theta = 70°$ at 1000 Hz in Fig. 1. Fig. 1(a) and (b) shows the marginal likelihoods of ITD and ILD, respectively, Fig. 1(c) shows two different DRR priors, and Fig. 1(d) and (e) shows the two resulting log-likelihood distributions with r marginalized. The joint log-likelihood functions in Fig. 1(d) and (e) are shown as equal contour plots, where (d) is generated using the descending prior (squares), and (e) is generated using the ascending prior (circles). $R = 5$ in this example. While each function exhibits two peaks, the primary peak in (e)

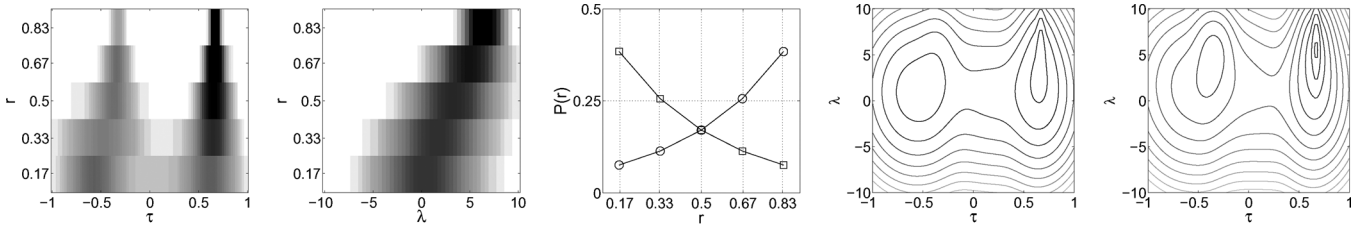

Fig. 1. Marginal ITD (a) and ILD (b) likelihoods, DRR prior (c), and equal contour plots of the ITD-ILD log-likelihood distributions (d) and (e) for $\theta = 70^\circ$ at 1000 Hz. The distribution in (d) uses the descending prior (squares) from (c), and the distribution in (e) uses the ascending prior (circles) from (c).

is much higher and sharper than the primary peak in (d) and is more selective in terms of ILD. Also note that the secondary peak in (d) has a slightly different ITD location and ILD much closer to 0 than the secondary peak in (e).

C. Model Training

Recent approaches to training binaural models of ITD and ILD incorporate simulations of multi-source pickup in a reverberant environment [27], [39], and thus may be sensitive to deviation from the room configuration or acoustic conditions used in training. In this work, we generate training mixtures by combining a point source with a simulated diffuse noise, and in doing so, avoid capturing environment-specific effects. We assume only the head-related transfer functions (HRTFs) of the binaural setup are known. We simulate a point source by filtering monaural signals using the HRTF for a given azimuth. The diffuse noise is created by passing uncorrelated noise signals through each of the HRTFs for the binaural setup and then adding them together. We provide more detail on the generation of training data in Section V-C.

Given a set of training data for a specific azimuth, we measure τ and λ from each pair of mixture T-F units and calculate r using (1). Since the simulated target includes only direct-path propagation, $x^E_{c,m}$ and $v^E_{c,m}$ are simply the premixed target and diffuse noise signals. We discretize the r values into R equally spaced bins. In this study we let $R = 5$ and have found the procedure to be relatively insensitive to the number of bins, provided a sufficient number (roughly 3 or more) is used. The total number of Gaussian components in the resulting model is proportional to R, thus choosing a small number limits the complexity of the model.

For each frequency channel, azimuth and DRR bin we learn the GMM parameters for the ITD dimension, $\{\rho_{c,k}(r,\tau_\theta), \mu_{c,k}(r,\tau_\theta), \sigma_{c,k}(r,\tau_\theta)\}$, using the EM algorithm, where $k \in \{1,\ldots,K_c\}$. We set the number of components, K_c, by determining the number of peaks in the range between -1.1 to 1.1 ms (to capture some edge effects) assuming that the cross-correlation function used to calculate ITD is periodic with the channel center frequency and that a peak exists at τ_θ. We then add one extra component to give the model more flexibility. The expected number of peaks in the cross-correlation function, and therefore K_c, increases systematically with center frequency. For each frequency channel, azimuth, and DRR bin we also measure the sample mean and variance for the ILD dimension, $\{\mu_c(r,\lambda_\theta), \sigma_c(r,\lambda_\theta)\}$. Finally, we calculate the number of data points that fall into each DRR bin for $P_c(r)$, although in order to remove the influence of training conditions, these values may be unused. We discuss how $P_c(r)$ is set for the experiments in this study in Section V-D.

III. Monaural Pathway

Both harmonicity and onset synchrony are known to be strong cues for across frequency grouping in auditory scene analysis [8] and have been shown to influence localization judgements by human listeners [4]. Motivated by this work and recent advances in monaural source segregation [37], the proposed framework incorporates a monaural pathway that uses a pitch-based and an onset/offset analysis to group T-F units dominated by the same underlying source. The grouping is used to constrain the integration of binaural cues for azimuth estimation.

We use existing algorithms for multipitch tracking [20] and onset/offset-based segmentation [17]. We also incorporate a pitch-based grouping method that is similar to the approach described in [19]. In this section, due to space constraints, we provide only a brief description of these methods and discuss their role in the proposed system. The interested reader is referred to the cited papers for more details.

A. Multipitch Tracking

In order to group T-F units based on pitch information, we incorporate a recent multipitch tracking system designed for reverberant and noisy speech [20]. This system estimates up to two pitch periods per time frame using a hidden Markov model (HMM) tracking framework. The state space of the HMM is a collection of subspaces corresponding to the cases with zero, one, or two voiced sources. The one- and two-source subspaces consist of all allowable single pitch and pitch combinations, respectively (covering the frequency range from 80 to 500 Hz). The model is allowed to jump between subspaces (i.e., the number of voiced sources can change), and pitch dynamics within a subspace are modeled by pitch transition probabilities. The observed data used in computation of state likelihoods is based on the correlogram [37]. The Viterbi algorithm is used to find the optimal path through the pitch subspaces, thereby estimating both the number of voiced sources and the corresponding pitch periods in each time frame.

We use this system to generate pitch estimates from both the left and right signals independently. Once pitch estimates are generated, we link pitch points across time when the change in pitch is below a predetermined threshold. We refer to a set of linked pitch points as a *pitch contour*. We use a threshold of 7% relative change in pitch frequency.

B. Pitch-Based Grouping

Pitch contours are used as the basis for grouping T-F units dominated by the same voiced source. For each individual pitch estimate, we use a supervised learning approach to identify T-F units across frequency that exhibit periodicity consistent with that of the estimate. Since the pitch estimates have already been linked across time intervals into pitch contours, T-F units associated with each pitch estimate are also grouped across time to form sets of T-F units, which we refer to as *simultaneous streams*.

Specifically, we use a multi-layer perceptron (MLP) to model the posterior probability that the dominant source in a T-F unit is consistent with a hypothesized pitch period. The features used as input to the MLP are extracted from the correlogram and envelope correlogram, calculated from both the left and right signals. The correlogram is a normalized running auto-correlation performed in each frequency channel for each time frame. We use a low-pass filter with 500-Hz cutoff frequency and a Kaiser window to extract signal envelopes. We train a separate MLP for each frequency channel, which consists of a hidden layer with 30 nodes. Training is accomplished using a generalized Levenberg–Marquardt backpropagation algorithm. We train the MLPs using a set of mixtures described in Section V-C. For each training mixture we extract the correlogram and envelope correlogram features, calculate the ideal binary mask (IBM) [37] and generate the *ground truth* pitch of the target signal by running the pitch estimation method proposed in [6] on the premixed signal. The IBM is used to provide the true classification label for each T-F unit and the ground truth pitch points are used to select the correlogram features corresponding to the pitch period of the target source. A more detailed description of models and training for pitch-based grouping can be found in [19].

C. Onset/Offset Based Segmentation

To capture unvoiced speech regions, the monaural pathway also incorporates the onset/offset segmentation approach proposed in [17]. The method first identifies onsets (increases in signal energy) and offsets (decreases in signal energy) across time within gammatone sub-bands. Detected onsets and offsets are linked across frequency into onset and offset fronts based on synchrony. Onset fronts are grouped with corresponding offset fronts based on frequency overlap. The set of T-F units between a pair of onset and offset fronts forms a T-F segment. Segmentation is performed with three different scales of across-time and across-frequency smoothing. Segments generated using the different smoothing scales are then integrated into a single set of T-F segments.

We use this segmentation system to generate T-F segments for the left and the right mixture independently. We make three changes to the implementation relative to that presented in [17]. First, to match the peripheral processing of the binaural pathway we implement the segmentation algorithm using 64 frequency channels, rather than 128. Second, we adjust the standard deviation of the Gaussian kernels used for across-frequency smoothing to account for the change from 128 to 64 channels. Third, in preliminary experiments we have found that pitch-based grouping is more reliable than the onset/offset segmentation in voiced speech regions. With this in mind, we eliminate T-F units from the segments if they are members of a pitch-based simultaneous stream.

D. Onset-Based Weights

In challenging acoustic environments, many T-F units will be corrupted by diffuse noise or reverberation. Although multiple aspects of the system seek to overcome this issue, we nevertheless find it beneficial to weight the contribution of T-F units to the localization decision so as to minimize the effect of units that are likely dominated by noise. Motivated by the precedence effect [23], we incorporate a simple cue weighting mechanism that identifies strong onsets in the mixture signal. We first extract the signal envelope for each frequency channel of the left and the right signal by squaring and passing each sub-band through a first-order IIR filter with a time constant of 10 ms. The resulting envelope signals are then decimated to a sample rate of 100 Hz (to match the frame rate of the other processing stages). Finally we compute

$$w_{c,m}^{E} = \left[\frac{e_c^E[m] - e_c^E[m-1]}{e_c^E[m-1]}\right]^{+} \tag{4}$$

as the weight for unit $u_{c,m}^{E}$. Here $e_c^E[m]$ denotes the sample of the decimated envelope signal corresponding to $u_{c,m}^{E}$ and $[\cdot]^{+}$ half-wave rectification.

IV. Localization Framework

The binaural pathway extracts azimuth-dependent information from each T-F unit pair while the monaural pathway groups T-F units that are likely to be dominated by the same source. The final stage of the proposed system then integrates this information and produces a set of N azimuth estimates. In [39], we developed a maximum likelihood framework for joint localization and labeling of pitch-based simultaneous streams. We take a similar approach here, but now deal with both voiced and unvoiced speech, and also use simultaneous streams and T-F segments generated from both the left and right mixture.

Conceptually speaking, to perform localization we first postulate a set of N possible azimuths, where we assume N is known. For each simultaneous stream or T-F segment we find the most likely azimuth from the postulated set and integrate likelihood scores over all streams and segments. The process generates a total likelihood for each postulated set of azimuths, and we choose the set that maximizes this value.

Formally, let I^E be the total number of simultaneous streams and T-F segments from ear signal E. Each individual simultaneous stream or T-F segment, denoted s_i^E, is a collection of T-F units. Assuming conditional independence between T-F units, the weighted log-likelihood for s_i^E is then

$$\beta_i^E(\theta) = \sum_{c,m \in s_i^E} w_{c,m}^{E} \ln\left(P_c(\tau_{c,m}, \lambda_{c,m}|\theta)\right). \tag{5}$$

We search for the most likely set of N azimuths using

$$\hat{\Theta} = \arg\max_{\Theta}\left(\sum_{i=1}^{I^L} \beta_i^L(\theta_{\hat{y}_i^L}) + \sum_{j=1}^{I^R} \beta_j^R(\theta_{\hat{y}_j^R})\right) \tag{6}$$

where $\Theta = \{\theta_0, \theta_1, \ldots, \theta_{N-1}\}$ denotes a set of N azimuths and

$$\hat{y}_i^E = \arg\max_{y \in \{0,1,\ldots,N-1\}} \beta_i^E(\theta_y). \tag{7}$$

V. Evaluation Methodology

We conduct three experiments to evaluate the effectiveness of the proposed method relative to existing systems. This section provides necessary details regarding the generation of training and evaluation data, and the binaural models, comparison systems and metrics used in the evaluation.

A. Binaural Impulse Responses

We use two different sets of binaural impulse responses (BIRs) in this study. One set is simulated and one set is measured in real environments. Each set assumes a different binaural setup, and we will refer to them according to the assumed setup.

The simulated BIRs, which we refer to as the KEMAR set, are generated using the ROOMSIM package [9]. This software combines the image method for reverberation [2] with HRTF measurements [15] made using a KEMAR dummy head. BIRs generated in this way represent a reasonable simulation of pickup by a KEMAR in real environments while allowing control of array and source placement, as well as characteristics of the room. We create a library of BIRs by generating ten room configurations, where room size, array position, and array orientation are set at random. We then generate BIRs for azimuths between -90° and 90°, spaced by 5°, at distances of 1, 2, and 4 m (where available in the room configuration). Reflection coefficients of the wall surfaces are set to be equal and to be the same across frequency, such that the reverberation time (T_{60}) is approximately 600 ms. In order to train binaural models, as described in Section II-C, we generate anechoic BIRs for the same azimuths using the HRTF measurements directly (i.e., no room simulation).

The other set includes publicly available measured BIRs, which are described in [18]. Impulse responses are measured using a head and torso simulator (HATS) in five different environments. Four environments are reverberant (rooms A, B, C, and D), with different sizes, reflective characteristics, and reverberation times. Measurements are also made in an anechoic environment. In all cases, BIRs are measured for azimuths between -90° and 90°, spaced by 5°, at a distance of 1.5 m. We use the BIRs from the three most reverberant rooms (B, C, and D) to generate an evaluation database, where the T_{60} times are listed as 0.47, 0.68, and 0.89 s, respectively. We use the anechoic measurements to train binaural models. We refer to this set of BIRs as the HATS set.

B. Evaluation Data

We create two evaluation sets, one from the KEMAR BIR set and one from the HATS BIR set. In the KEMAR evaluation set we consider 2 or 3 target talkers, source distances of 1, 2, and 4 m, and infinite, 6 and 0 dB speech-to-noise ratios (SNRs) for a total of 18 conditions. We generate 100 binaural mixtures for each condition. Azimuths are selected randomly such that sources are spaced by 10° or more. The SNR is set by summing the energy of all speech sources relative to a simulated diffuse noise. The energy of both left and right channels is summed prior to SNR calculation. Speech sources are simulated by filtering monaural utterances, drawn randomly from the TIMIT database [16], by a selected KEMAR BIR. Monaural utterances, originally sampled at 16 kHz, are upsampled to 44.1 kHz to match the rate of the KEMAR BIRs. The diffuse noise is created by filtering uncorrelated speech-shaped noise signals through each of the anechoic KEMAR BIRs and then adding them together. We create the speech-shaped filter by averaging the amplitude spectra of 200 speech utterances drawn from TIMIT at random. Each mixture has a length of 2 s, where monaural speech utterances are concatenated so that they are sufficiently long (if needed). We employ an energy threshold to eliminate silence at the beginning and end of the monaural utterances in order to ensure that speech sources are active in the majority of time frames.

We create the HATS evaluation set in the same way. In this case we consider 2 target talkers in three rooms (B, C and D), and infinite, 6 and 0 dB SNRs, giving us a total of nine conditions. All other details are as described for the KEMAR set.

C. Training Data

To train binaural models we generate data using the anechoic KEMAR and HATS BIRs. For each BIR set we generate 250 speech plus noise mixtures per azimuth where, as described in Section II-C, we simulate anechoic speech using a BIR for a selected azimuth and simulate diffuse speech-shaped noise as described in Section V-B. Speech utterances are drawn randomly from TIMIT. The only factors varying between mixtures are the speech utterances used and the input SNR, which is selected randomly to be -24, -12, -6, -3, 0, 3, 6, 12, or 24 dB.

In order to evaluate how well the proposed scheme for training binaural models compares to a more ideal training scenario, we also generate a training set using the reverberant HATS BIRs. We generate 250 mixtures for each azimuth and for each of the three rooms seen in the HATS evaluation set. The procedure used to generate these training mixtures is identical to that used for the evaluation mixtures, however, each training mixture generated for a specific azimuth contains one speech source placed at that azimuth.

Finally, we generate a set of 100 mixtures to train the MLPs used for pitch-based grouping. Each mixture contains a dominant speech source corrupted by a multi-talker babble consisting of 10, 15, or 20 interfering speech sources. Monaural speech utterances are drawn randomly from the TIMIT database and filtered by a selected KEMAR BIR. The azimuth of all sources is selected randomly between -90° and 90° and the SNR between the dominant talker and the multi-talker interference is set at random between -6 and 12 dB (in 3 dB steps).

D. Binaural Models

Using the training procedure outlined in Section II-C along with the anechoic speech plus diffuse noise data described in the previous subsection, we create KEMAR and HATS models. In addition to using the HATS models trained from anechoic measurements, we generate a set of models for the HATS evaluation set that we refer to as *matched*. The matched models are created using the second set of training mixtures described in the previous subsection. A separate model is trained for each room. In this case, target signals are simulated by convolution with a measured, reverberant impulse response. It is therefore necessary to

approximate direct-path propagation of the target in order to calculate the DRR. To accomplish this we identify the approximate location of the direct-path component by finding the largest peak in the BIR, then truncate the impulse response 10 ms after the start of the direct-path component. For the HATS BIRs used in this study, we have found that choosing 10 ms ensures capture of the full direct-path component, while minimizing the number of reflections included. This parameter may vary for different measurements, but is not necessary to train models based on measurements made in a controlled environment.

The choice of values for the DRR prior, $P_c(r)$, will influence the shape of the resulting likelihood distribution (see Fig. 1). If $P_c(r)$ is set empirically (i.e., by counting the number of training data points that fall into each DRR bin), the distributions will reflect the acoustic conditions seen in training. If one desires to minimize the influence of training data, $P_c(r)$, can be set according to some assumptions about the acoustics that will be seen in practice. As described in Section II-C, we discretize DRR into 5 bins, corresponding to values of 0.83, 0.67, 0.5, 0.33, and 0.17, or approximately 7, 3, 0, -3 and -7 in dB. For the KEMAR and HATS models, we set $P_c(0.17) = 0.6$, $P_c(0.33) = 0.1$, $P_c(0.5) = 0.1$, $P_c(0.67) = 0.1$, $P_c(0.83) = 0.1$ for all frequencies and azimuths. We chose these values to inject limited knowledge of the evaluation set acoustics. Specifically, this prior reflects an assumption that a given T-F unit is more likely to be dominated by the residual signal (noise or reverberation) than the direct-path of a speech source. These specific values were chosen by an informal analysis of a small number of mixtures that resemble those seen in the evaluation set. Since the matched models for the HATS evaluation set are trained using data that perfectly matches the conditions that will be seen in testing, we set $P_c(r)$ empirically for the matched models.

E. Comparison Systems

In the experiments below, we compare performance of the proposed method with two existing methods from the literature [11], [26]. The system proposed in [11], denoted SRP, is a steered beamformer that incorporates the phase transform weighting to increase robustness in reverberant conditions. Our implementation measures the response power over 20-ms time frames that overlap by 50%. We integrate over frequencies up to 8 kHz, since the TIMIT sources do not have energy beyond this frequency, sum the responses across time and select the N most prominent peaks as the source azimuths. We consider the same set of azimuths used in the proposed method and use the direct-path interaural phase differences of the KEMAR or HATS array, depending on the evaluation set, for beam steering.

The second comparison system used is the joint localization and segregation approach presented in [26], dubbed MESSL, and is representative of the spatial clustering approach to localization. We use an implementation of MESSL provided by the algorithm's author. The system requires specification of the number of sources and iteratively fits GMM models of interaural phase difference (IPD) and ILD to the observed data using an EM procedure. Across frequency integration is handled by tying GMM models in individual frequency bands to a principal ITD. Based on the model fits, we find the most likely ITD for each source and map this to an azimuth estimate using the group delay of the anechoic KEMAR or HATS BIRs, depending on the evaluation set. MESSL is initialized using the PHAT-Histogram method [1], where we use the group delay of the anechoic KEMAR or HATS BIRs to specify the ITD bins for the histogram. Mixture signals are first downsampled from 44.1 to 16 kHz because the original TIMIT sources were sampled at 16 kHz.

We selected these methods from a set of candidates that also included the systems proposed in [1], [24], [40]. We found that in most conditions, the performance of MESSL and PHAT-Histogram [1] was comparable, but that MESSL outperformed PHAT-Histogram for short integration times. We also found the stencil filter method in [24] to perform similarly, but systematically worse than the SRP method. Finally, we found the clustering method proposed in [40] to perform poorly on our data set. The system was unable to localize sources at angles more lateral than 45° even in single-source anechoic conditions, due to the large number of frequencies in which spatial aliasing was present.

F. Evaluation Metrics

In all cases and for all methods we assume oracle knowledge of the number of speech sources. With this knowledge we seek to estimate the azimuth angle of each source based on a fixed amount of observed data. We evaluate the different localization systems using two metrics. For each evaluation mixture, we consider a source to be detected if there is an azimuth estimate within (and including) 10°. We then measure the *gross accuracy* as the percentage of detected sources. We also measure the average azimuth error of those estimates that were within the 10° threshold and refer to this as the *fine error*. Note that a single azimuth estimate cannot be used to detect more than one source.

VI. Evaluation Results

In this section, we present the results from three experiments. The first experiment analyzes the impact of monaural cues on localization. The second experiment provides a comparison of the proposed method to existing systems using simulated impulse responses. The third experiment tests generalization of the system to measured impulse responses and robustness using mismatched binaural models.

A. Experiment 1: Influence of Monaural Cues

In this experiment, we analyze the influence of monaural cues on localization performance. We compare performance to two binaural baselines that use the proposed azimuth-dependent models but do not incorporate monaural cues. The first baseline, denoted Binaural-Hist, uses the procedure proposed in [27]. This approach estimates the dominant azimuth in each frame according to

$$\hat{\theta}_m = \arg\max_{\theta} \sum_c P_c(\tau_{c,m}, \lambda_{c,m}|\theta) \tag{8}$$

then generates an across-time histogram of the frame-level azimuth estimates and selects the N largest histogram peaks as the source azimuths. The second baseline method, denoted Binaural-ML, is a maximum likelihood procedure similar to the

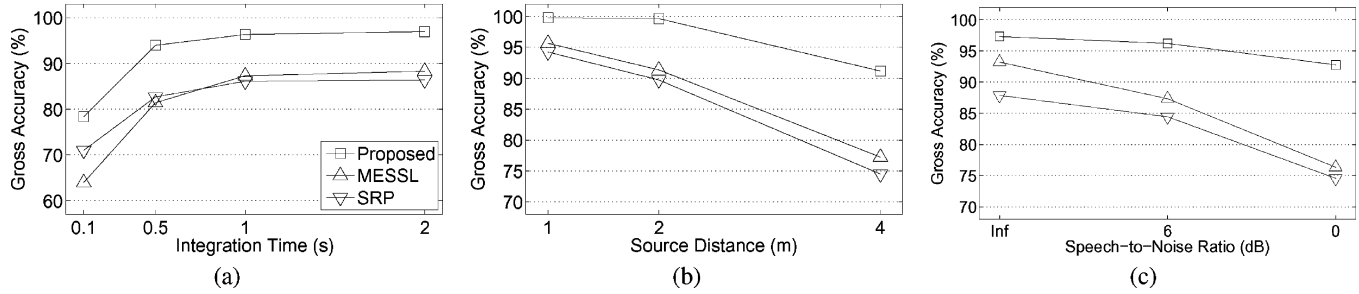

Fig. 2. Gross accuracy (%) shown over the two-talker KEMAR set as a function of (a) integration time, (b) distance, and (c) noise level. In (b) and (c), we show results for a 2-s integration time. The legend in (a) is applicable to all figures shown.

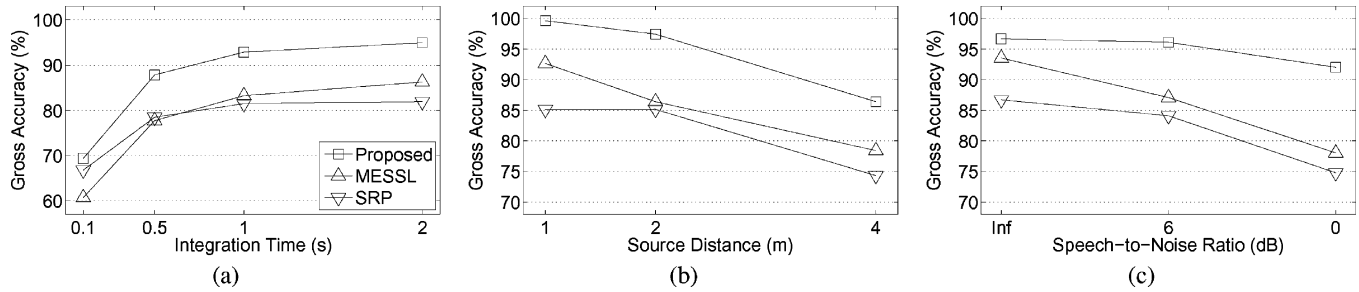

Fig. 3. Gross accuracy (%) shown over the three-talker KEMAR set as a function of (a) integration time, (b) distance, and (c) noise level. In (b) and (c), we show results for a 2-s integration time. The legend in (a) is applicable to all figures shown.

TABLE I
GROSS ACCURACY (%) FOR THE KEMAR SET FOR ALTERNATIVE T-F INTEGRATION METHODS

	Two-talker	Three-talker
Binaural-Hist	90.3	84.1
Binaural-ML	91.7	86.9
Binaural + Pitch-based Grouping	96.2	93.1
Binaural + Onset/offset Segmentation	95.6	92.3
Proposed	97.0	94.6

proposed method, but does not incorporate monaural grouping. In this case, azimuth estimates are derived using

$$\hat{\Theta} = \arg\max_{\Theta} \sum_{c,m} \ln\left(P_c(\tau_{c,m}, \lambda_{c,m}|\theta_{\hat{y}_{c,m}})\right) \quad (9)$$

$$\hat{y}_{c,m} = \arg\max_{y\in\{0,\ldots,N-1\}} P_c(\tau_{c,m}, \lambda_{c,m}|\theta_y) \quad (10)$$

where $\Theta = \{\theta_0, \ldots, \theta_{N-1}\}$ is a set of N azimuths and $\hat{y}_{c,m}$ is an integer to specify the assignment of T-F unit $u_{c,m}$ to one of the sources. The Binaural-ML system performs segregation on the basis of binaural cues, similar to [26], [30], and derives each azimuth estimate from a subset of T-F units. Along with the binaural baselines, we evaluate three variations of the proposed system, where we consider only pitch-based grouping, only onset/offset segmentation, and the full proposed system. Performance differences between the two baselines and different variations of the proposed system are entirely due to how binaural information is integrated across time and frequency.

Table I shows the gross accuracy over the entire set of two- and three-talker KEMAR mixtures. We first note that that the Binaural-ML method provides a small improvement over the Binaural-Hist approach. This gain can be attributed to the fact that evidence for multiple sources can be extracted from even a single time frame, which is not possible with the Binaural-Hist approach. However, the rather marginal gain suggests that while it is conceptually appealing to perform joint segregation and localization, there appears to be little improvement in localization when the segregation process is based entirely on binaural cues. In contrast, all systems that incorporate monaural grouping achieve substantial gains relative to the binaural baselines. The best performance is achieved by the full system that incorporates both types of monaural grouping and onset-based weights, where we see a nearly 8% absolute gain in gross accuracy relative to the Binaural-ML approach on the three-talker mixtures.

We also note that in addition to the constraints enforced on T-F grouping, the monaural mechanisms select a subset of the T-F units for binaural integration. On the KEMAR data set, about 84% of T-F units are selected. The number of talkers and the source distance appear to have a very small influence on this percentage, while decreasing the SNR can substantially reduce the percentage of T-F units selected. On average, the percentage of T-F units selected decreases from roughly 91% at infinite SNR to 79% at 0 dB SNR.

B. Experiment 2: Comparison on Kemar Evaluation Set

In this experiment, we compare localization performance of the proposed system to the two comparison methods from the literature [11], [26] on the KEMAR evaluation set. We present the gross accuracy for various experimental conditions in Figs. 2 and 3. We show results considering integration times of 0.1, 0.5, 1, and 2 s in Figs. 2(a) and 3(a). We do so by providing each system the mixture signals from beginning to the specified time. Results for different integration times are averaged over all distances and SNRs. We show results as a function of source distance in Figs. 2(b) and 3(b). In this case, we generate results using the entire mixture (2 s) and average results over SNRs.

TABLE II
GROSS ACCURACY (%) AND FINE ERROR (°) FOR THE KEMAR SET

	Gross Accuracy		Fine Error	
	Two-talker	Three-talker	Two-talker	Three-talker
Proposed	97.0	94.6	1.0	1.3
MESSL	88.5	85.6	1.5	1.9
SRP	86.6	81.6	1.4	1.8

Similarly, we show results as a function of SNR in Figs. 2(c) and 3(c) using the entire mixture and average over source distances. As one would expect, all systems perform better as more data is used for the estimate, while there is a systematic decrease in performance as sources become more distant or the background noise level increases.

We can see that the proposed system outperforms the comparison methods in terms of gross accuracy for all evaluation conditions. MESSL outperforms SRP when the integration time is 1 s or longer. On the shortest integration time, 0.1 s, the initialization of MESSL by PHAT-Histogram [1] is poor, and the algorithm is more likely to have large errors than SRP. The improvement in gross accuracy by the proposed system over MESSL is 8.8% (absolute), calculated over the entire two- and three-talker evaluation set. The improvement in gross accuracy relative to SRP is 11.7% over the entire evaluation set. In Figs. 2(b), (c), 3(b), and (c), we see that the improvement achieved by the proposed system tends to be larger in the more difficult conditions with distant sources and strong background noise. For example, on the two-talker evaluation set with sources at 4 m and 0-dB SNR, the improvement in gross accuracy is about 23% relative to both MESSL and SRP.

In Table II we show the gross accuracy and the fine error on the full two- and three-talker data sets when using a 2-s integration time. As previously stated, the gross accuracy using the proposed method is higher than for the comparison methods on both the two- and three-talker data and we can also see that the fine error is lower. Since the proposed system utilizes prior training, the performance increase relative to comparison methods is due to both the inclusion of monaural cues and the prior knowledge captured by the binaural model. Although there are numerous differences between the Binaural-ML system (see Section VI-A) and the comparison methods, some indication of the relative contribution of monaural cues and the binaural model can be gained by noting that the Binaural-ML system achieves a 2.3% and 5.2% gain in gross accuracy relative to MESSL and SRP, respectively, while the proposed method achieves the 8.8% and 11.7% gains noted above.

To test the necessity of prior training with HRTFs of the binaural setup that will be seen in testing, we also performed tests with binaural models trained on HRTFs that simulate microphone pickup on the surface of a rigid sphere [14]. We found degradation in terms of gross accuracy to be only 3.4% and 4.5% on the two- and three-talker data sets, respectively. Degradation in terms of fine error was larger, from 1.0° with the KEMAR models to 3.3° with the sphere-based models on the two-talker set, and from 1.3° to 3.1° on the three-talker set. These results indicate that the proposed method can still perform well even with no prior knowledge of the binaural setup to be used in practice.

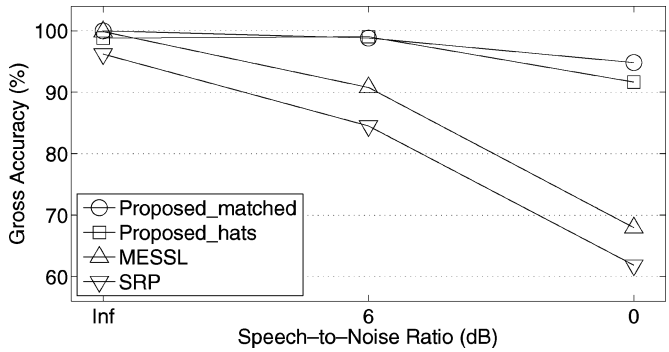

Fig. 4. Gross accuracy (%) as a function of noise level for the HATS evaluation set with an integration time of 2 s.

As one might expect from studies of localization acuity in human subjects [5], the azimuth error is lower near the median plane than to the side of the head when using the proposed method. Across the entire two-talker data set, the average error (error over all estimates, not the fine error) for sources with azimuth between −30° and 30° is 0.6°, whereas it increases to 4° for sources with azimuth more lateral than 60°. We also note that gross accuracy is lower in test cases where sources are spaced more closely.

C. Experiment 3: Hats Evaluation Set

In this experiment, we compare localization performance of the proposed system to the two comparison methods on the HATS evaluation set, which uses measured BIRs from real room environments. We also compare the performance achieved using the HATS models trained on anechoic measurements to the matched models trained on the BIRs seen in testing. We assume that using the matched models will provide a performance upper bound and are interested in the amount of degradation due to using mismatched models. Performance using the HATS models on this evaluation set should give the best indication of how the system would perform in a practical setting where calibration measurements may be assumed, but extensive training in real environments would not be available.

We present the gross accuracy as a function of SNR in Fig. 4, where results are averaged over all rooms and an integration time of 2 s is used. Notable is the fact that the difference in gross accuracy between the matched models and the HATS models is 1.1% or less for the infinite and 6-dB mixtures and 3.2% for the 0-dB mixtures. Consistent with Experiment 2, the performance improvement achieved by the proposed system relative to the comparison methods increases as the level of background noise increases.

In Table III, we show the gross accuracy and the fine error for all four systems on each room in the HATS set separately, with a 2-s integration time. We see that the HATS models perform comparably to the matched models in terms of gross accuracy, and the proposed system with HATS models achieves a gross accuracy about 10% higher than MESSL and about 15% higher than SRP. However, we can see that the fine error is consistently lower when using the matched models. The fine error is similar for all three realizable systems, with MESSL achieving

TABLE III
GROSS ACCURACY (%) AND FINE ERROR (°) FOR THE HATS SET

	Gross Accuracy			Fine Error		
	B	C	D	B	C	D
Proposed_matched	98.7	97.4	97.6	0.6	0.6	0.4
Proposed_HATS	97.3	96.9	95.4	1.1	0.8	1.8
MESSL	87.6	86.8	84.2	1.0	1.1	1.0
SRP	82.2	80.5	79.8	1.0	1.0	1.8

the lowest fine error on average. The larger fine error for the proposed system with HATS model and the SRP system on the Room D data is due to a systematic discrepancy between the direct-path cues of the anechoic measurements and the direct-path cues of the Room D measurements.

VII. CONCLUDING REMARKS

The results in Section VI clearly demonstrate the effectiveness of the proposed localization method. By integrating monaural CASA methods with an azimuth-dependent model of ITD and ILD, we are able to accurately localize multiple sources in adverse conditions. The method yields a significant improvement over baseline methods that do not incorporate monaural grouping. The results from Experiment 1 support the perspective that monaural segregation can facilitate localization. The results from Experiment 2 show that localization improvement is largest in adverse conditions and for distant sources and the results from Experiment 3 establish the robustness of the proposed method when using impulse responses measured in real room environments.

We have also proposed a flexible binaural model that can be easily adapted to different binaural setups and acoustic conditions. Results from Experiments 2 and 3 indicate that robust performance can be achieved with only anechoic measurements of the binaural setup, and thus the simulations used to train the models proposed in [27] and [39] may be unnecessary. Although only briefly discussed here, preliminary results for generalization to unseen binaural setups are promising.

Since we generate pitch-based simultaneous streams and onset/offset based segments from both the left and right signals, some of the resulting sets of T-F units will overlap in time and frequency, thus the independence assumption made in order to derive (6) and (7) is clearly violated. Considering dependencies between simultaneous streams and T-F segments will increase computational complexity of the system; however, it is possible that doing so could improve performance.

In this work, we assumed prior knowledge of the number of sources, and thus a key problem for future work is estimating the number of sources. The computational complexity of the search over azimuth combinations limits the total number of sources that can be detected by the proposed method to a small number (e.g., 3 or 4), but as the proposed method is a maximum likelihood approach, several well-known model selection criteria such as Akaike information criterion or minimum description length could be employed in (6) to penalize overestimating the number of sources. Our preliminary analysis with such a penalty term has produced promising results. We also note that in cases when the number of sources is misestimated or wrongly provided to the proposed system, we observe that our system still produces reasonable results. If the number of sources is underestimated, the system tends to localize the most dominant sources, while if the number is overestimated, the system tends to break one azimuth into two closely spaced ones.

Another important extension of the proposed system is to the case with a time-varying number of moving sources. The results presented in this study suggest that incorporating monaural cues improves assignment of T-F units to source signals, and as such, monaural cues could potentially benefit detection and tracking of moving sources. This is a topic that will be addressed in future work.

ACKNOWLEDGMENT

The authors would like to thank the three anonymous reviewers for their constructive criticisms and suggestions. The authors would also like to thank M. Mandel and T. May for making implementations of their algorithms available, and C. Hummersone for making the set of measured impulse responses available.

REFERENCES

[1] P. Aarabi, "Self-localization dynamic microphone arrays," *IEEE Trans. Syst., Man, Cybern. C*, vol. 32, no. 4, pp. 474–484, Nov. 2002.
[2] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, pp. 943–950, 1979.
[3] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *J. Acoust. Soc. Amer.*, vol. 107, no. 5, pp. 384–391, 2000.
[4] V. Best, F. J. Gallun, S. Carlile, and B. G. Shinn-Cunningham, "Binaural interference and auditory grouping," *J. Acoust. Soc. Amer.*, vol. 121, no. 2, pp. 1070–1076, 2007.
[5] J. Blauert, *Spatial Hearing – The Psychophysics of Human Sound Localization*. Cambridge, MA: MIT Press, 1997.
[6] P. Boersma, "Accurate short-time analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," *Inst. Phon. Sci.*, vol. 17, pp. 97–110, 1993.
[7] M. Brandstein, "Time-delay estimation of reverberated speech exploiting harmonic structure," *J. Acoust. Soc. Amer.*, vol. 105, pp. 2914–2919, 1999.
[8] A. S. Bregman, *Auditory Scene Analysis*. Cambridge, MA: MIT Press, 1990.
[9] D. R. Campbell, The ROOMSIM User Guide (v3.3) 2004 [Online]. Available: http://media.paisley.ac.uk/~campbell/Roomsim/
[10] H. Christensen, N. Ma, S. N. Wrigley, and J. Barker, "A speech fragment approach to localizing multiple speakers in reverberant environments," in *Proc. ICASSP*, Apr. 2009, pp. 4593–4596.
[11] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brstein and D. Ward, Eds. New York: Springer, 2001, ch. 8, pp. 157–180.
[12] M. Dietz, S. D. Ewert, and V. Hohmann, "Auditory model based direction estimation of concurrent speakers from binaural signals," *Speech Commun.*, vol. 53, pp. 592–605, 2011.
[13] S. Doclo and M. Moonen, "Robust adaptive time delay estimation for speaker localization in noisy and reverberant acoustic environments," *EURASIP J. App. Signal Process.*, vol. 2003, pp. 1110–1124, 2003.
[14] R. O. Duda and W. L. Martens, "Range dependence of the response of a spherical head model," *J. Acoust. Soc. Amer.*, vol. 104, no. 5, pp. 3048–3058, 1998.
[15] W. G. Gardner and K. D. Martin, "HRTF measurements of a KEMAR," *J. Acoust. Soc. Amer.*, vol. 97, pp. 3907–3908, 1995.
[16] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus," 1993 [Online]. Available: http://www.ldc.upenn.edu/Catalog/LDC93S1.html
[17] G. Hu and D. L. Wang, "Auditory segmentation based on onset and offset analysis," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 15, no. 2, pp. 396–405, Feb. 2007.

[18] C. Hummersone, R. Mason, and T. Brookes, "Dynamic precedence effect modeling for source separation in reverberant environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1867–1871, Sep. 2010.
[19] Z. Jin and D. L. Wang, "A supervised learning approach to monaural segregation of reverberant speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 625–638, May 2009.
[20] Z. Jin and D. L. Wang, "HMM-based multipitch tracking for noisy and reverberant speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 5, pp. 1091–1102, Jul. 2011.
[21] M. Képesi, F. Pernkopf, and M. Wohlmayr, "Joint position pitch tracking for 2-channel audio," in *Proc. Int. Workshop Content Based Multimedia Indexing*, 2007.
[22] 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.
[23] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman, "The precedence effect," *J. Acoust. Soc. Amer.*, vol. 106, pp. 1633–1654, 1999.
[24] C. Liu, B. C. Wheeler, W. D. O'Brien, R. C. Bilger, C. R. Lansing, and A. S. Feng, "Localization of multiple sound sources with two microphones," *J. Acoust. Soc. Amer.*, vol. 108, pp. 1888–1905, 2000.
[25] W.-K. Ma, B.-N. Vo, S. Singh, and A. Baddelay, "Tracking an unknown time-varying number of speakers using TDOA measurements: A random finite set approach," *IEEE Trans. Signal Process.*, vol. 54, no. 9, pp. 3291–3304, Sep. 2006.
[26] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 382–394, Feb. 2010.
[27] T. May, S. van de Par, and A. Kohlrausch, "A probabilistic model for robust localization based on a binaural auditory frond-end," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 1–13, Jan. 2011.
[28] L. Y. Ngan, Y. Wu, C. So, P. C. Ching, and S. W. Lee, "Joint time delay and pitch estimation for speaker localization," in *Proc. ICAS*, 2003.
[29] J. Nix and V. Hohmann, "Combined estimation of spectral envelopes and sound source direction of concurrent voices by multidimensional statistical filtering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 995–1008, Mar. 2007.
[30] Ö. Yılmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.
[31] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," Tech. Rep. MRC App. Psych. Unit, Cambridge, MA, 1988.
[32] M. Raspaud, H. Viste, and G. Evangelista, "Binaural source localization by joint estimation of ILD and ITD," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 1, pp. 68–77, Jan. 2010.
[33] N. Roman and D. L. Wang, "Binaural tracking of multiple moving sources," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 4, pp. 728–739, May 2008.
[34] N. Roman, D. L. Wang, and G. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Amer.*, vol. 114, pp. 2236–2252, 2003.
[35] A. Stéphenne and B. Champagne, "A new cepstral prefiltering technique for estimating time delay under reverberant conditions," *Signal Process.*, vol. 59, no. 3, pp. 253–266, 1997.
[36] R. M. Stern, G. J. Brown, and D. L. Wang, "Binaural sound localization," in *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, D. L. Wang and G. J. Brown, Eds. New York: Wiley, 2006, pp. 147–185.
[37] D. L. Wang and G. J. Brown, Eds., Computational Auditory Scene Analysis: Principles, Algorithms, and Applications Hoboken, NJ, Wiley/IEEE Press, 2006.
[38] J. Woodruff and D. L. Wang, "Integrating monaural and binaural analysis for localizing multiple reverberant sound sources," in *Proc. ICASSP*, Mar. 2010, pp. 2706–2709.
[39] J. Woodruff and D. L. Wang, "Sequential organization of speech in reverberant environments by integrating monaural grouping and binaural localization," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 18, no. 7, pp. 1856–1866, Sep. 2010.
[40] W. Zhang and B. D. Rao, "A two microphone-based approach for source localization of multiple speech sources," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 8, pp. 1913–1928, Dec. 2010.

John Woodruff (S'09) received the B.F.A. degree in performing arts and technology and the B.S. degree in mathematics from the University of Michigan, Ann Arbor, in 2002 and 2004, respectively, and the M.Mus. degree in music technology from Northwestern University, Evanston, IL, in 2006. He is currently pursuing the Ph.D. degree in computer science and engineering at The Ohio State University, Columbus.

His research interests include computational auditory scene analysis, music and speech processing, auditory perception, and statistical learning.

DeLiang Wang, (F'04) photograph and biography not available at the time of publication.