Sequential Organization of Speech in Reverberant Environments by Integrating Monaural Grouping and Binaural Localization

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

***Abstract*—Existing binaural approaches to speech segregation place an exclusive burden on cues related to the location of sound sources in space. These approaches can achieve excellent performance in anechoic conditions but degrade rapidly in realistic environments where room reverberation corrupts localization cues. In this paper, we propose to integrate monaural and binaural processing to achieve segregation and localization of voiced speech in reverberant environments. The proposed approach builds on monaural analysis for simultaneous organization, and combines it with a novel method for generation of location-based cues in a probabilistic framework that jointly achieves localization and sequential organization. We compare localization performance to two existing methods, sequential organization performance to a model-based system that uses only monaural cues, and segregation performance to an exclusively binaural system. Results suggest that the proposed framework allows for improved source localization and robust segregation of voiced speech in environments with considerable reverberation.**

***Index Terms*—Binaural speech segregation, computational auditory scene analysis, monaural grouping, sequential organization, sound localization.**

I. Introduction

MOST existing approaches to binaural or sensor-array-based speech segregation have relied exclusively on localization cues embedded in the differences between signals recorded by multiple microphones [1], [2]. These approaches may be characterized as spatial filtering (or beamforming), which enhances the signal from a specific direction. Spatial filtering approaches can be very effective in certain acoustic conditions. On the other hand, beamforming has well-known limitations. Chief among them is substantial performance degradation in reverberant environments. Rigid surfaces reflect a sound source incident upon them, hence corrupting localization cues [3].

Manuscript received October 04, 2009; revised May 03, 2010. Date of current version August 13, 2010. This work was supported 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 in part by a grant from the Oticon Foundation. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Tomohiro Nakatani.

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.2010.2050087

Time–frequency masking techniques have been proposed to deal with segregation in reverberant environments [4], [5]. Recent approaches have relied on probabilistic frameworks that jointly perform source localization and time–frequency masking to segregate multiple sources [6]–[8]. These approaches improve segregation by modeling the increased variability of localization cues in reverberation, and improve localization by integrating cues over part of the mixture in which a given source is dominant. In spite of the performance gain achieved by such systems, they are still fundamentally limited by the discriminative power of localization cues, which is substantially diminished in environments with room reverberation.

In this paper, we propose an alternative framework that integrates monaural and binaural analysis to achieve robust localization and segregation of voiced speech in reverberant environments. In the language of *auditory scene analysis* (ASA) [9], our proposed system uses monaural cues to achieve *simultaneous organization*, or grouping sound components of the mixture across frequency and short, continuous time intervals. This allows locally extracted, unreliable binaural cues to be integrated over large time–frequency regions. Integration over such regions enhances localization robustness in reverberant conditions and in turn, we use robust localization to achieve *sequential organization*, or grouping sound components of the mixture across disparate intervals of time.

Our computational framework is partly motivated by psychoacoustic studies suggesting that binaural cues may not play a dominant role in simultaneous organization, but are important for sequential organization [10], [11]. Further, human listeners are able to effectively localize multiple sound sources in reverberant environments [12], and some recent analysis suggests that localization may be facilitated by monaural grouping, rather than localization acting as a fundamental grouping cue in ASA [13].

Prior work exploring the integration of monaural and binaural cues for reverberant speech processing is limited. In [14], localization cues are used to perform initial segregation in reverberant conditions. Initial segregation provides a favorable starting point for estimating the pitch track of the target voice, which is then used to further enhance the target signal. In [15], pitch and ITD are used to achieve localization of simultaneous speakers in reverberant environments. Our prior work analyzed the impact of idealized monaural grouping on localization and segregation of speech in reverberant environments [16], and showed that pitch-based monaural grouping can be used to

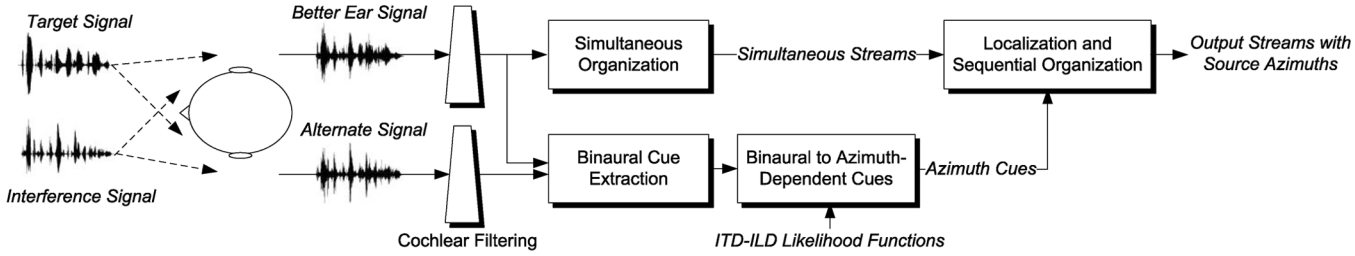


Fig. 1. Schematic diagram of the proposed system. Binaural recordings are fed as input to the system. Cochlear filtering is applied to both the left and right ear signals. Monaural processing generates simultaneous streams from the *Better Ear Signal*. Both signals are used to generate azimuth-dependent cues. Simultaneous streams and azimuth-dependent cues are combined in the final localization and sequential organization stage.

achieve accurate localization of multiple sources in noisy and reverberant environments [17].

Utilizing binaural cues to handle sequential organization is attractive because monaural features alone may not be able to solve the problem. For example, in a mixture of two male speakers who have a similar vocal range, pitch-based features cannot be used to group components of the mixture that are far apart in time. As a result, feature-based monaural systems have largely avoided sequential organization by focusing on short utterances of voiced speech [18] or assuming prior knowledge of the target signal's pitch [19], or achieved sequential organization by assuming speech mixed with non-speech interference [20].

Shao and Wang explicitly addressed sequential organization in a monaural system using a model-based approach [21]. They use feature-based monaural processing to perform simultaneous organization of voiced speech, and speaker identification to perform sequential organization of the already formed time–frequency segments. They provide extensive results on sequential organization performance in co-channel speech mixtures as well as speech mixed with non-speech intrusions. However, they do not address sequential organization in reverberant environments.

In the following section, we provide an overview of the proposed architecture. In Section III we discuss monaural simultaneous organization of voiced speech. Section IV outlines our methods for extraction of binaural cues, for calculating azimuth-dependent cues, and a mechanism for weighting cues based on their expected reliability. In Section V, we formulate joint sequential organization and localization in a probabilistic framework. We assess both localization and sequential organization performance, and compare the proposed system to existing methods in Section VI. We conclude with a discussion in Section VII.

II. System Overview

The proposed system integrates monaural and binaural analysis to achieve segregation of voiced speech. A diagram is provided in Fig. 1. The input to the system is a binaural recording of a speech source mixed with one or more interfering signals. The recordings are assumed to be made with two microphones inserted in the ear canals of a human listener or dummy head, and we will refer to the two mixture signals as the left ear and right ear signals, denoted by $l(n)$ and $r(n)$, respectively.

In this paper, we use the ROOMSIM package [22] to generate impulse responses that simulate binaural input at human ears. This package uses measured *head-related transfer function* (HRTF) data from a KEMAR dummy head [23] in combination with the image method for simulating room acoustics [24]. To generate binaural speech mixtures, we use monaural speech signals drawn from the TIMIT database [25], pass the signals through a binaural impulse response pair, and sum the resulting binaural target and interference signals to create a binaural mixture. The TIMIT signals, originally sampled at 16 kHz, are upsampled to 44.1 kHz prior to binaural filtering to match the sampling rate of the impulse responses.

When processing a given mixture, the system first passes both the left and right signals through a bank of 128 gammatone filters [26] with center frequencies from 50 to 8000 Hz spaced on the equivalent rectangular bandwidth scale [27]. Since the source signals are originally sampled at 16 kHz, the filterbank covers the entire frequency range of speech energy in the mixtures. We denote the signals for frequency channel c as $l_c(n)$ and $r_c(n)$. Each filtered signal is divided into 20-ms time frames with a frame shift of 10 ms to create a *cochleagram* [2] of time–frequency (T-F) units for both the left and right ear signals. A T-F unit, which we denote as $u_{c,m}$, is an elemental sound component that contains one frame of signal, indexed by m, from one of the gammatone filter outputs, indexed by c.

In the first stage of the system, the tandem algorithm of Hu and Wang [28], [29] is used to form *simultaneous streams* from the T-F units of the *better ear* signal. By better ear signal, we mean the signal in which the input SNR is higher, as determined from the signals before mixing. A simultaneous stream refers to a collection of T-F units over a continuous time interval that are thought to be dominated by the same source. A *stream*, in the computational auditory scene analysis (CASA) literature, typically corresponds to the set of T-F units dominated by a specific source. A simultaneous stream refers to a continuous part of a stream that is grouped through simultaneous organization (i.e., through across frequency grouping and temporal continuity). The tandem algorithm generates simultaneous streams for voiced speech using monaural cues such as harmonicity and amplitude modulation. Unvoiced speech presents a greater challenge for monaural systems and is not dealt with in this study (see [20]).

Binaural cues are extracted that measure differences in timing and level between corresponding T-F units of the left and right ear signals. A set of trained, azimuth-dependent likelihood functions are then used to map from timing and level differences to cues related to source location. Azimuth cues are integrated within simultaneous streams in a probabilistic framework to

achieve sequential organization and to estimate the underlying source locations. The output of the system is a set of streams, one for each source in the mixture, and the azimuth angles of the underlying sources.

III. Simultaneous Organization

Simultaneous organization in CASA systems forms simultaneous streams, each of which may contain disconnected T-F segments across frequency but span a continuous time interval. We use the tandem algorithm proposed in [28], [29] to generate simultaneous streams for voiced regions of the better ear mixture. The tandem algorithm iteratively estimates a set of pitch contours and associated simultaneous streams. In a first pass, T-F segments that contain voiced speech are identified using cross-channel correlation of correlogram responses. Up to two pitch points per time frame are estimated by finding peaks in the summary correlogram created from only the selected, voiced T-F segments. For each pitch point found, T-F units that are consistent with that pitch are identified using a set of trained multi-layer perceptrons (one for each frequency channel). Pitch points and associated sets of T-F units are linked across continuous time intervals to form pitch contours and associated simultaneous streams using a criterion that measures pitch deviation and spectral continuity. Pitch contours and simultaneous streams that span only a single time frame are discarded. Finally, the pitch contours and associated simultaneous streams are iteratively refined until convergence.

We focus on multi-talker mixtures in reverberant environments, and find that in this case the criterion used in the tandem algorithm for connecting pitch points and simultaneous streams across continuous time intervals is too liberal. For this reason, we break pitch contours and simultaneous streams when the pitch deviation between time frames is large. Specifically, let τ_1 and τ_2 be pitch periods from the same contour in neighboring time frames. If $|\log_2(\tau_1/\tau_2)| > 0.08$, the contour and associated simultaneous streams are broken into two contours and two simultaneous streams. The value of 0.08 was selected on the basis of informal analysis, and was not specifically tuned for optimal performance on the data set discussed in Section VI.

An example set of pitch contours and simultaneous streams are shown in Fig. 2. The plots are generated using the better ear mixture of a female talker placed at -15° azimuth and a male talker placed at 30° azimuth in a reverberant environment with 0.4 s reverberation time (T_{60}). There are a total of 27 contour and simultaneous stream pairs shown. The energy of each T-F unit in the cochleagram of the mixture is shown in Fig. 2(a). In Fig. 2(b), detected pitch contours are shown by alternating between circles and squares, while ground truth pitch points generated from the reverberant signals prior to mixing are shown as solid lines. In Fig. 2(c), each gray level corresponds to a separate simultaneous stream. One can see that simultaneous streams may contain multiple segments across frequency but are continuous in time.

IV. Binaural Processing

In this section, we describe how binaural cues are extracted from the mixture signals and propose a mechanism to translate these cues into information about the azimuth of the underlying source signals. We also discuss a method to weight binaural cues according to their expected reliability.

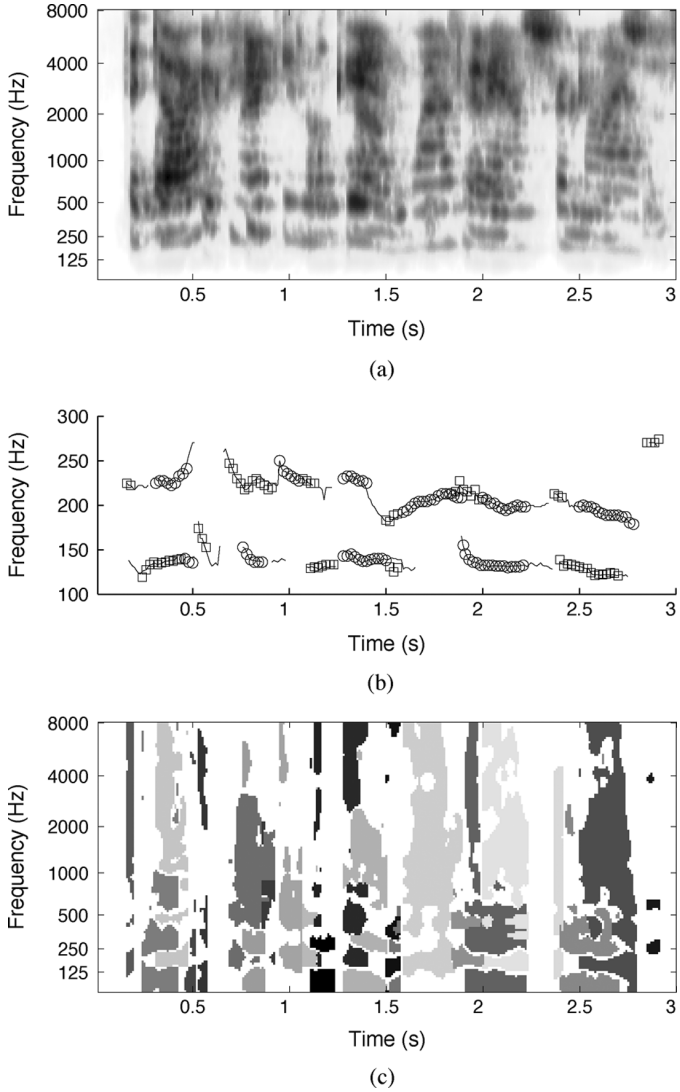


Fig. 2. Example of multi-pitch detection and simultaneous organization using the tandem algorithm. (a) Cochleagram of a two talker mixture. (b) Ground truth pitch points (solid lines) and detected pitches (circles and squares). Different pitch contours are shown by alternating between circles and squares. (c) Simultaneous streams corresponding to different pitch contours are shown with different gray levels.

A. Binaural Cue Extraction

Two primary binaural cues used by humans for localization of sound sources are interaural time difference (ITD) and interaural level difference (ILD) [30]. We calculate ITD in individual frequency bands by first computing the normalized cross-correlation,

$$C(c,m,\tau) = \frac{\sum_{n=0}^{T_n-1} l_c\left(m\frac{T_n}{2}-n\right) r_c\left(m\frac{T_n}{2}-n-\tau\right)}{\sqrt{\sum_{n=0}^{T_n-1}\left(l_c\left(m\frac{T_n}{2}-n\right)\right)^2}\sqrt{\sum_{n=0}^{T_n-1}\left(r_c\left(m\frac{T_n}{2}-n-\tau\right)\right)^2}} \quad (1)$$

where τ is the time lag for the correlation and $\tau \in [-44, 44]$, c and m index frequency channels and time frames, respectively,

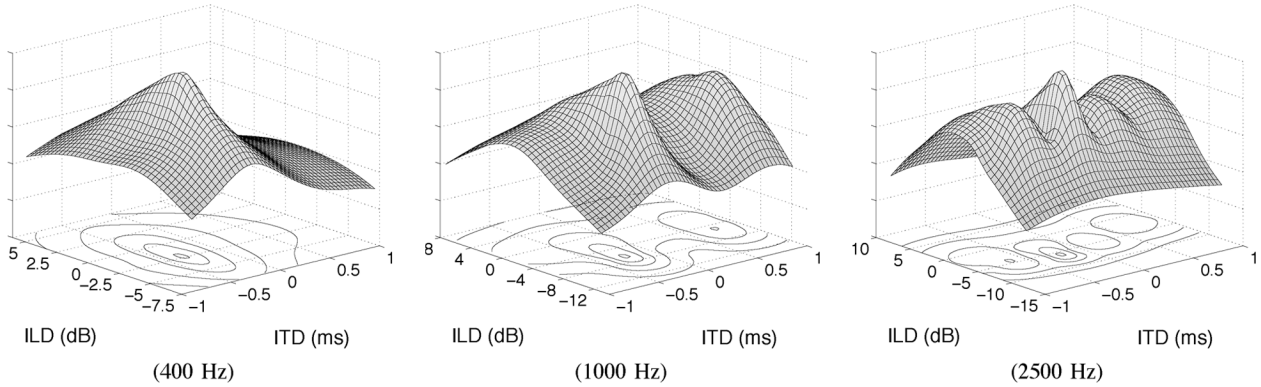


Fig. 3. Examples of ITD-ILD likelihood functions for azimuth 25° at frequencies of 400, 1000, and 2500 Hz. Each example shows the log-likelihood as a surface with projected contour plots that show cross sections of the function at equally spaced intervals.

T_n denotes the number of samples per time frame and the frame shift is $(T_n/2)$. The ITD is then defined as the time lag that produces the maximum peak in the normalized cross-correlation function, or

$$\tau_{c,m} = \arg\max_{\tau \in L} C(c, m, \tau) \tag{2}$$

where L denotes the set of peak lags in $C(c, m, \tau)$.

ILD corresponds to the energy ratio in dB between the two signals in corresponding T-F units:

$$\lambda_{c,m} = 10 \log_{10} \left(\frac{\sum_{n=0}^{T_n - 1} \left(l_c \left(m \frac{T_n}{2} - n \right) \right)^2}{\sum_{n=0}^{T_n - 1} \left(r_c \left(m \frac{T_n}{2} - n \right) \right)^2} \right). \tag{3}$$

B. Azimuth-Dependent Likelihood Functions

If one assumes binaural sensors in an anechoic environment, a given source position relative to the listener's ears will produce a specific, frequency dependent set of ITDs and ILDs for that listener. In order to effectively integrate information across frequency for a given source position, these patterns must be taken into account. Further, integration of ITD and ILD cues extracted from reverberant mixtures of multiple sources should account for deviations from the free-field patterns.

In this paper, we focus on a subset of possible source locations. Specifically, we restrict the sources to be in front of the listener with 0° elevation. As a result, source localization reduces to azimuth estimation in the interval $[-90°, 90°]$. To translate from raw ITD-ILD information to azimuth, we train a joint ITD-ILD likelihood function, $P_c(\tau_{c,m}, \lambda_{c,m}|\theta)$, for each azimuth, θ, and frequency channel, c. Likelihood functions are trained on single-source speech in various room configurations and reverberation conditions using kernel density estimation [31]. The room size, listener position, source distance to listener and reflection coefficients of the wall surfaces are randomly selected from a predefined set of 540 possibilities. Following Roman *et al.*, we use Gaussian kernels for density estimation and choose smoothing parameters using the least-squares cross-validation method [31]. For a more detailed description, see [32].

An ITD-ILD likelihood function is generated for each of 37 azimuths, $[-90°, 90°]$ spaced by 5°, and for each of the 128 frequency channels. With these functions, we can translate the ITD-ILD values measured from a given T-F unit pair into an azimuth-dependent likelihood curve. Due to reverberation, we do not expect the maximum of the likelihood curve for each T-F unit pair to be a good indication of the dominant source's azimuth, but hope that a good indication of the dominant source's azimuth emerges through integration over a simultaneous stream.

The set of likelihood distributions for a specific azimuth captures the frequency dependent pattern of ITDs and ILDs for that azimuth and the multi-peak ambiguities present at higher frequencies where signal wavelengths are shorter than the distance between ears or microphones. Each distribution has a peak corresponding to the free-field cues for that angle, but also captures common deviations from the free-field cues due to reverberation. We show three distributions in Fig. 3 for azimuth 25°. Note that, in addition to the above points, the azimuth-dependent distributions capture the complementary nature of localization cues [30] in that ITD provides greater discrimination between angles at lower frequencies (note the large ILD variation in the 400 Hz example) and ILD improves discrimination between angles at higher frequencies where spatial aliasing hinders discrimination by ITD alone.

Our approach is adapted from the one proposed in [32]. In that system two ITD-ILD likelihood functions are trained for each frequency channel, $P_c(\tau_{c,m}, \lambda_{c,m}|H_0)$ and $P_c(\tau_{c,m}, \lambda_{c,m}|H_1)$, where H_0 denotes the hypothesis that the target signal is stronger than the interference signal, and H_1 that the target is weaker. The distributions $P_c(\tau_{c,m}, \lambda_{c,m}|H_0)$ and $P_c(\tau_{c,m}, \lambda_{c,m}|H_1)$ are trained for each target/interference angle configuration. The ITD search space is limited around the expected free-field target ITD in both training and testing to avoid the multi-peak ambiguity in higher frequency channels. For a test utterance, the azimuths of both target and interference sources are estimated, the appropriate set of likelihood distributions is selected and the maximum *a posteriori* decision rule is used to estimate a binary mask for the target source.

There are two primary reasons for altering the method in [32] to the one proposed here. First, our proposed approach lowers the training burden because likelihood functions are trained for each angle individually, rather than as combinations of angles. Second, the fact that we do not limit the ITD search space in training allows us to use the likelihood functions in estimation

of the underlying source azimuths, rather than requiring a preliminary stage to estimate the angles. In [17], we showed that our proposed localization method, which utilizes the ITD-ILD likelihood functions, performs significantly better than the method proposed in [32].

Because we do not limit the ITD search space, our approach does not attempt to resolve the multi-peak ambiguity inherent in high frequency ITD calculation at the T-F unit level. For frequency channels in which the wavelength of the signal is shorter than the spacing between microphones, multiple peaks are captured by the likelihood functions (see Fig. 3). Spatial aliasing in these channels is naturally resolved by integrating across frequency within a simultaneous stream.

C. Cue Weighting

In reverberant recordings, many T-F units will contain cues that differ significantly from free-field cues. Although these deviations are incorporated in the training of the ITD-ILD likelihood functions described above, including a weighting function or cue selection mechanism that indicates when an azimuth cue should be trusted can improve localization performance. Motivated by the *precedence effect* [33], we incorporate a simple cue weighting mechanism that identifies strong onsets in the mixture signal. When a large increase in energy occurs, and shortly thereafter, the azimuth cues are expected to be more reliable. We therefore generate a weight $w_{c,m}$ associated with $u_{c,m}$ that measures the change in signal energy over time. First, we define a recursive method to measure the average signal energy in both left and right channels as follows:

$$e_c(n) = \alpha(l_c(n)^2 + r_c(n)^2) + (1-\alpha)e_c(n-1). \quad (4)$$

Here $\alpha \in [0,1]$ and $\alpha = (1)/(Wf_s)$, where W denotes the time constant for integration and f_s is the sampling frequency of the signals. We set $W = 10$ ms and $f_s = 44.1$ kHz. We then calculate the percent of change in energy between samples and average over an integration window to get

$$w_{c,m} = \frac{1}{T_n}\sum_{n=0}^{T_n-1}\frac{e_c\left(m\frac{T_n}{2}-n\right) - e_c\left(m\frac{T_n}{2}-n-1\right)}{e_c\left(m\frac{T_n}{2}-n-1\right)}. \quad (5)$$

$w_{c,m}$ is then normalized over each mixture to have values between 0 and 1 by first subtracting the minimum value over all T-F units, finding the maximum value after subtraction, and then dividing by the maximum value over all T-F units.

We have found measuring change in energy using this method to provide better results than simply taking the change in average energy from unit to unit, or taking the more traditional derivative of the signal envelope [2]. We have also found better performance by keeping only those weights above a specified threshold. The difficulty with a fixed threshold however, is that one may end up with a simultaneous stream with no unit above the threshold. To avoid this we set a threshold for each simultaneous stream so that the set of T-F units exceeding the threshold retain 25% of the signal energy in the simultaneous stream. $w_{c,m}$ is set to 0 for all T-F units below the selected threshold. We have found that the system is not particularly sensitive to the value of 25% and that values between about 15% and 40% give similar performance in terms of localization accuracy.

Alternative selection mechanisms have been proposed in the literature [34], [35], [15]. Faller and Merimaa proposed *interaural coherence* as a cue selection mechanism [34], although in preliminary experiments we found the proposed method to outperform selection methods based on interaural coherence. The method proposed in [35] uses ridge regression to learn a finite-impulse response filter that predicts localization precision for single-source reverberant speech in stationary noise. This method essentially identifies strong signal onsets, as does our approach, but requires training. The study in [15] finds that a precedence motivated cue weighting scheme performs about as well as two alternatives on a database of two-talker mixtures in a small office environment.

V. Localization and Sequential Organization

As described above, the first stage of the system generates simultaneous streams for voiced regions of the better ear mixture and extracts azimuth-dependent cues for all T-F units using the left and right ear mixtures. In this section, we describe the source localization and sequential organization process. The goal of sequential organization is to generate a target or interference label for each of the simultaneous streams, thereby grouping the simultaneous streams that occur mainly at different times. Our approach jointly determines the source azimuths and sequential organization (simultaneous stream labeling) that maximizes the likelihood of the binaural data. This approach is inspired by the model-based sequential organization scheme proposed in [36].

Let N be the number of sources in the mixture, and I be the number of simultaneous streams formed using monaural analysis. Denote the set of all possible azimuths as Θ and the set of simultaneous streams as $S = \{s_1, s_2, \ldots, s_I\}$. Let Y be the set of all N^I sequential organizations, or labelings, of the set S and y be a specific organization. We seek to maximize the joint probability of a set of angles and a sequential organization given the observed data, D. This can be expressed as

$$\hat{\theta}_0, \ldots, \hat{\theta}_{N-1}, \hat{y} = \underset{\theta_0,\ldots,\theta_{N-1}\in\Theta, y\in Y}{\arg\max} P(\theta_0, \ldots, \theta_{N-1}, y|D). \quad (6)$$

For simplicity, assume that $N = 2$ and apply Bayes rule to get

$$\begin{aligned}\hat{\theta}_0, \hat{\theta}_1, \hat{y} &= \underset{\theta_0,\theta_1\in\Theta,\theta_0\neq\theta_1,y\in Y}{\arg\max} \frac{P(D|\theta_0,\theta_1,y)P(\theta_0,\theta_1,y)}{P(D)} \\ &= \underset{\theta_0,\theta_1\in\Theta,\theta_0\neq\theta_1,y\in Y}{\arg\max} P(D|\theta_0,\theta_1,y) \end{aligned} \quad (7)$$

assuming that all angles and sequential organizations are equally likely (with the exception that $P(\theta_0 = \theta_1) = 0$).

Now, let S_0 be the set of simultaneous streams associated with θ_0 and S_1 be the set of simultaneous streams associated with θ_1 by y. Using ITD and ILD as the observed mixture data, and assuming independence between simultaneous streams and

between T-F units of the same simultaneous stream, we can express (7) as

$$\hat{\theta}_0, \hat{\theta}_1, \hat{y} = \underset{\theta_0,\theta_1\in\Theta,\theta_0\neq\theta_1,y\in Y}{\arg\max} \left(\prod_{s_i\in S_0} \prod_{u_{c,m}\in s_i} P_c(\tau_{c,m}, \lambda_{c,m}|\theta_0) \cdot \prod_{s_j\in S_1} \prod_{u_{c,m}\in s_j} P_c(\tau_{c,m}, \lambda_{c,m}|\theta_1) \right) \quad (8)$$

where P_c denotes a probability function defined for frequency channel c (see Section IV-B).

One can express the above equation as two separate equations that can be solved simultaneously in one polynomial-time operation as

$$\hat{y}_i = \underset{y_i\in\{0,1\}}{\arg\max} \left(\sum_{u_{c,m}\in s_i} \log(P_c(\tau_{c,m}, \lambda_{c,m}|\theta_{y_i})) \right) \quad (9)$$

$$\hat{\theta}_0, \hat{\theta}_1 = \underset{\theta_0,\theta_1\in\Theta,\theta_0\neq\theta_1}{\arg\max} \left(\sum_{i=1}^{I} \sum_{u_{c,m}\in s_i} \log(P_c(\tau_{c,m}, \lambda_{c,m}|\theta_{\hat{y}_i})) \right) \quad (10)$$

where $\hat{y}_i$ denotes the label assigned to s_i. The key assumption in moving to (9) and (10) is the independence between simultaneous streams expressed in (8).

Incorporating the weighting parameter defined in Section IV-C, (9) and (10) become

$$\hat{y}_i = \underset{y_i\in\{0,1\}}{\arg\max} \left(\sum_{u_{c,m}\in s_i} w_{c,m} \log(P_c(\tau_{c,m}, \lambda_{c,m}|\theta_{y_i})) \right) \quad (11)$$

$$\hat{\theta}_0, \hat{\theta}_1 = \underset{\theta_0,\theta_1\in\Theta,\theta_0\neq\theta_1}{\arg\max} \left(\sum_{i=1}^{I} \sum_{u_{c,m}\in s_i} w_{c,m} \log(P_c(\tau_{c,m}, \lambda_{c,m}|\theta_{\hat{y}_i})) \right). \quad (12)$$

For the case with $N > 2$, use $y_i \in \{0, 1, \ldots, N-1\}$ rather than $y_i \in \{0, 1\}$ in (11) and $\{\theta_0, \theta_1, \ldots, \theta_{N-1}\} \in \Theta, \theta_i \neq \theta_j$ in (12). The complexity of the search space is $I(\binom{|\Theta|}{N})$, which is reasonable when the number of sources of interest is relatively small and the size of the azimuth space is moderate. In our experiments in Section VI, $|\Theta| = 37$ and $N \leq 3$. We provide a more thorough discussion regarding search complexity and independence assumptions in Section VII.

VI. Evaluation and Comparison

In this section, we evaluate source localization, localization-based sequential organization, and segregation of voiced speech using the proposed integration of monaural and binaural processing. We analyze localization performance with and without the cue weighting mechanism discussed in Section IV-C and compare the proposed method to two existing methods in various reverberation conditions. We evaluate sequential organization performance in various reverberation conditions through comparison to a model-based approach and to a method that incorporates prior knowledge. Finally, we evaluate voiced speech segregation of the full system through comparison to an exclusively binaural approach and to identify the conditions in which integration of monaural and binaural analysis can outperform binaural analysis alone.

A. Training and Mixture Generation

We generate a training and a testing library of binaural impulse responses for 37 direct sound azimuths between -90° and 90° spaced by 5°, and 7 T_{60} times between 0 and 0.8 s using the ROOMSIM package [22]. In the training library, three room size configurations, three source distances from the listener (0.5, 1 and 1.5 m) and five listener positions in the room are used. In the testing library, two room size configurations (different from those in training), three source distances from the listener (same as those in training) and two listener positions (different from those in training) are used. For training the ITD-ILD likelihood distributions, speech signals randomly selected from the eight dialect regions in the training portion of the TIMIT database [25] are upsampled to 44.1 kHz and convolved with a randomly selected impulse response pair from the training library (for a specified angle). Training is performed over 100 reverberant signals for each of the 37 azimuths (see Section IV-B).

For all testing mixtures we select target and interference speech signals from the TIMIT database, upsample the signals to 44.1 kHz, pass the signals through an impulse response pair from the testing library for a desired azimuth and T_{60} time, and sum the resulting binaural target and interference signals to create a binaural mixture. We generate 200 two-talker mixtures and 200 three-talker mixtures for each of the reverberation conditions. In each mixture the room dimensions, source distance and listener position are randomly selected and applied to all sources. For the two-talker mixtures source azimuths are selected randomly to be between 10° and 125° apart. For the three-talker mixtures source azimuths are selected randomly to be at least 10° apart. The average azimuth spacing over each set of two-talker mixtures is 53°, whereas the average spacing from the target source to the closest interference source is 41° for each set of three-talker mixtures. Speech utterances, azimuths and room conditions remain constant across different T_{60} times. Only the reflection coefficient of the wall surfaces was changed to achieve the selected T_{60}. The SNR of each mixture is set to 0 dB using the dry, monaural TIMIT utterances. This results in better ear mixtures that average 2.8 dB in anechoic conditions down to 1 dB in 0.8 s T_{60} for the two-talker case, and -0.4 dB in the anechoic mixtures down to -1.6 dB in 0.8 s T_{60} for the three-talker case. Mixture lengths are determined using the target utterance with the interference signals either truncated or concatenated with themselves to match the target length. In order to make a comparison to the model-based approach (discussed further in Section VI-C), the speakers used for the test mixtures are drawn from the set of 38 speakers in the DR1 dialect region of the TIMIT training database.

B. Localization Performance

In this section, we analyze the localization accuracy of the method described in Section V. Specifically, we measure average azimuth estimation error with and without cue weighting. We also compare localization performance to two existing methods for localization of multiple sound sources, as proposed

in [37] and [38], and to an exclusively binaural system that incorporates the azimuth-dependent likelihood functions described in Section IV-B, but labels each T-F unit independently.

The approach proposed by Liu *et al.* in [38], termed the *stencil filter*, performs coincidence detection for each frequency bin and time frame and counts the detected ITD as evidence for a particular azimuth if it falls along the azimuth's "primary" or "secondary" traces. The primary trace is simply the predicted ITD for that angle, while the secondary traces are due to ambiguity at higher frequencies. For comparison on the database described, some changes were necessary to account for the (somewhat) frequency-dependent nature of ITDs as detected by a binaural system and the discrete azimuth space. Further, because angles are assumed constant over the length of the mixture, azimuth responses from the stencil filter were integrated over all time frames for added accuracy and the two most prominent peaks were selected as the underlying source angles.

The SRP-PHAT algorithm is a well-known technique for localization in reverberant conditions [37]. It combines a steered beamformer with the phase transform weighting of the generalized cross-correlation. Rather than use a frequency-independent time delay to steer the beam pattern, as is typically done, we use the true frequency dependent phase delays of the anechoic HRTFs for each of the 37 possible angles. This resulted in much better performance across all conditions, and this information was also used for the stencil filter implementation. We measure the PHAT weighted steered response power for each angle over time frames of 1024 samples, or 23 ms, that overlap by 50% and integrate over frequencies up to 8 kHz, since the speech sources in our test corpus do not have energy beyond this frequency. We integrate over all time frames and again select the two most prominent peaks as the underlying source angles.

The exclusively binaural system treats each T-F unit independently and jointly estimates source azimuths and time–frequency masks. Specifically, for a given set of angle hypotheses $\{\hat{\theta}_0, \ldots, \hat{\theta}_{N-1}\}$, each T-F unit is given a source assignment $y_{c,m}$ using the azimuth-dependent likelihood functions. The azimuth set that maximizes the likelihood after integration over all T-F units is selected. This can be expressed with a slight alteration of (9) and (10)

$$\hat{y}_{c,m} = \underset{y_{c,m} \in \{0,\ldots,N-1\}}{\arg\max} P_c(\tau_{c,m}, \lambda_{c,m} | \theta_{y_{c,m}}) \quad (13)$$

$$\hat{\theta}_0, \ldots, \hat{\theta}_{N-1} = \underset{\theta_0,\ldots,\theta_{N-1} \in \Theta}{\arg\max} \sum_{u_{c,m}} P_c(\tau_{c,m}, \lambda_{c,m} | \theta_{\hat{y}_{c,m}}). \quad (14)$$

This approach is similar in spirit to [6] and [7] in that source azimuths and time–frequency masks are jointly estimated, allowing localization cues to be integrated over a subset of T-F units in the mixture. One key difference is that the binaural system presented here takes advantage of the pretrained, nonparametric likelihood functions whereas [6] and [7] fit parametric models directly to the observed mixture. It is important to note that we do not incorporate the voiced simultaneous streams in any way, thus unlike the proposed system, the binaural localization system makes use of both voiced and unvoiced speech.

Average azimuth error on the two-talker mixtures is shown in Fig. 4. Estimation is performed for 400 source signals (2 in each of the 200 two-talker mixtures) and for 7 T_{60} times. The results indicate that including weights associated with signal onsets improves azimuth estimation of the proposed method when significant reverberation is present. We can also see that both proposed methods outperform the existing methods for T_{60} of 300 ms or larger. The improvement relative to the stencil filter method averages 5.18° over the T_{60} range of 400 ms to 800 ms, 3.74° relative to the SRP-PHAT approach, and 3.51° relative to the exclusively binaural approach.

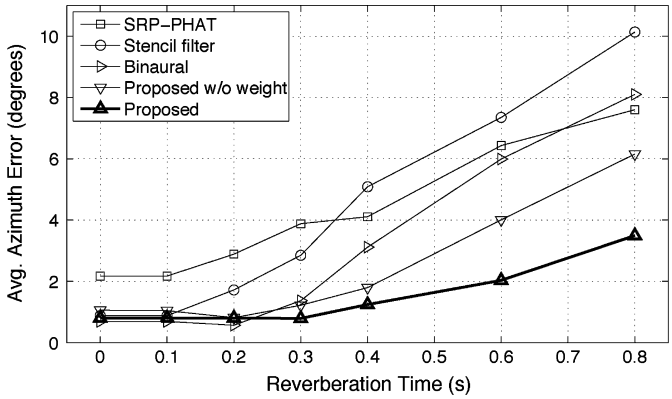


Fig. 4. Azimuth estimation error averaged over 200 two-talker mixtures, or 400 utterances, for various reverberation times. Results are shown using the proposed approach with and without cue weighting, and three alternative approaches.

The difference in performance between the methods is largely captured by how well they localize *both* sources in the mixtures. If we consider only the source that was localized with the most precision, the average azimuth error of all methods was near or below 2° in all T_{60} times. However, the proposed method was able to localize the second source with far more accuracy than the alternative methods. When T_{60} ranges from 400 ms to 800 ms, the proposed method decreased the average azimuth error of the less accurately localized source by between 60% and 70% relative to the alternative systems.

Performance on the three-talker mixtures followed the same trends, with the proposed system providing an accuracy improvement of 33%, 41%, and 48% over the binaural, SRP-PHAT and stencil filter methods, respectively, over the T_{60} range of 300 ms to 800 ms. The proposed system achieved about 5° azimuth error on this set of reverberant mixtures, averaged over the 600 sources (3 in each of the 200 mixtures) localized in each of the 4 T_{60} times.

The key advantage of both the proposed system and the binaural system is that azimuth-dependent cues for a particular source are not integrated over the entire mixture, as they are in the stencil filter and SRP-PHAT approaches. The comparison between the proposed method without cue weighting and the binaural method shows that monaural grouping alone facilitates more accurate localization as T-F units are not treated completely independent of one another. Selecting a subset of the T-F units using a mechanism for cue weighting is also advantageous in terms of localization accuracy.

C. Sequential Organization and Segregation Performance

To analyze both sequential organization and voiced segregation performance we use the *ideal binary mask* (IBM), which has been proposed as a main computational goal of CASA [39]

and shown to dramatically improve speech intelligibility when applied to noisy mixtures [40]. The IBM is a binary labeling of mixture T-F units such that when target energy is stronger than interference energy, the T-F unit is labeled with 1, and when target energy is weaker, the T-F unit is labeled with 0. Note that the IBM labels not only T-F units corresponding to voiced speech, but also those corresponding to unvoiced speech. We evaluate performance by finding the percentage of mixture energy contained in the simultaneous streams that is correctly labeled by an estimated mask, where ground truth labeling of a T-F unit in a simultaneous stream is generated using the IBM of the better ear mixture. We refer to this metric as the labeling accuracy. We measure the mixture energy in dB.

To evaluate sequential organization, we compare performance against a "ceiling" measure that incorporates ideal knowledge and to a recent model-based system [21]. We refer to the ceiling performance measure as *ideal sequential organization* (Ideal S.O.). In this case, a target/interference decision is made for each simultaneous stream based on whether the majority of the mixture energy is labeled target or interference by the IBM.

The model-based system uses pretrained speaker models to perform sequential organization of simultaneous streams for voiced speech [21]. Speaker models are trained using an auditory feature, gammatone frequency cepstral coefficients [21], and the system incorporates missing data reconstruction and uncertainty decoding to handle simultaneous streams that do not cover the full frequency range. The system is designed for anechoic speech trained in matched acoustic conditions. To account for both the azimuth-dependent HRTF filtering and reverberation contained in the mixture signals used in our database, some adjustments were made. First, we train speaker models for each of the reverberation conditions that will be seen in testing. For each of the 38 speakers, we select seven out of ten utterances for training, generate ten variations of each of these utterances with randomly selected azimuths for each of the seven reverberation times. This helps to minimize the mismatch between training and testing conditions, although as mentioned above, the impulse responses used in training are different from those in testing. We found this approach to give better performance than feature compensation methods (e.g., cepstral mean subtraction, and cepstral mean subtraction and variance normalization) for mismatched training and testing conditions.

In [21], a background model is used to allow the system to process speech mixed with multiple speech intrusions or nonspeech intrusions. Since we focus on the two and three-talker cases, we found that assuming all speakers are known *a priori* produces better results than using a generic background model. Incorporating this prior knowledge ensures that we are comparing to a high level of performance potentially achievable by the model-based system.

To identify the conditions in which the proposed integration of monaural and binaural analysis can improve segregation relative to binaural analysis alone, we compare performance to the exclusively binaural system described in (13) and (14). For the purpose of comparison, we still measure the labeling accuracy *within* the simultaneous streams, even though the exclusively binaural approach is able to generate a binary mask for the entire mixture.

As previously stated, the exclusively binaural system has much in common with the systems proposed in [6] and [7]. The key difference is that the binaural system presented here uses pre-trained, non-parametric likelihood functions rather than fitting parametric models to the observed mixture. To test whether models that are tuned to capture the reverberation condition of a specific mixture improves performance, we trained alternative non-parametric likelihood functions tuned for each T_{60} time of the test database. On our two-talker database we found little benefit in using the T_{60}-specific models for either the exclusively binaural or the proposed system (0.3% better on average for both systems). In training the likelihood functions as described in Section VI-A, we have generated a binaural model that, while specific to the binaural microphone (or listener) used for training, provides good performance across a variety of room conditions.

In Fig. 5, we show the performance of the proposed system, the model-based system, the binaural system and the ideal sequential organization scheme on the two- and three-talker mixtures. The performance achieved by Ideal S.O. indicates the quality of the monaural simultaneous organization. Any decrease below 100% reflects that the simultaneous streams are not exclusively dominated by target or interference. On the two-talker mixtures shown in Fig. 5(a), labeling error due to monaural analysis averages 11.6% across all T_{60} times, and is largely consistent across reverberation conditions. The performance difference between Ideal S.O. and the model-based or proposed systems reflects errors due to sequential organization. Model-based sequential organization introduces an additional 12.7% labeling error, averaged over all T_{60} times. The error introduced by localization-based sequential organization ranges from 1.8% in low reverberation conditions, up to almost 8% in the most reverberant condition. The relative performance improvement over the model-based system ranges between 9.5% and 14%, depending on the T_{60} time. This is notable, especially considering that the model-based results incorporate prior knowledge of the speaker identities contained in the mixture and the T_{60} time of the mixture. The proposed system outperforms the model-based approach on the three-talker mixtures as well [see Fig. 5(b)], although the gap is not as large.

In comparing the proposed system to the Ideal S.O. system, one can see that the proportion of labeling error attributable to localization-based sequential organization increases with both T_{60} time and the number of talkers, suggesting that an increase in the number of talkers or the reverberation time has a larger impact on the binaural sequential organization than on the accuracy of the monaural grouping. However, since all results are obtained from voiced speech only, as generated from the tandem algorithm's simultaneous streams, these measures do not penalize the simultaneous organization stage for what one might call misses, or T-F units that contain primarily voiced energy from one of the source signals, but are not captured by any of the simultaneous streams. We note that the proportion of total mixture energy (both voiced and unvoiced) that is captured by a simultaneous stream is 57% in the two-talker anechoic case, decreases to 35% averaged over the two-talker mixtures between 300 ms and 800 ms T_{60} and 33% averaged over the three-talker

(From the authors: Please note that Fig. 5(a) and 5(b) have been mistakenly reversed.)

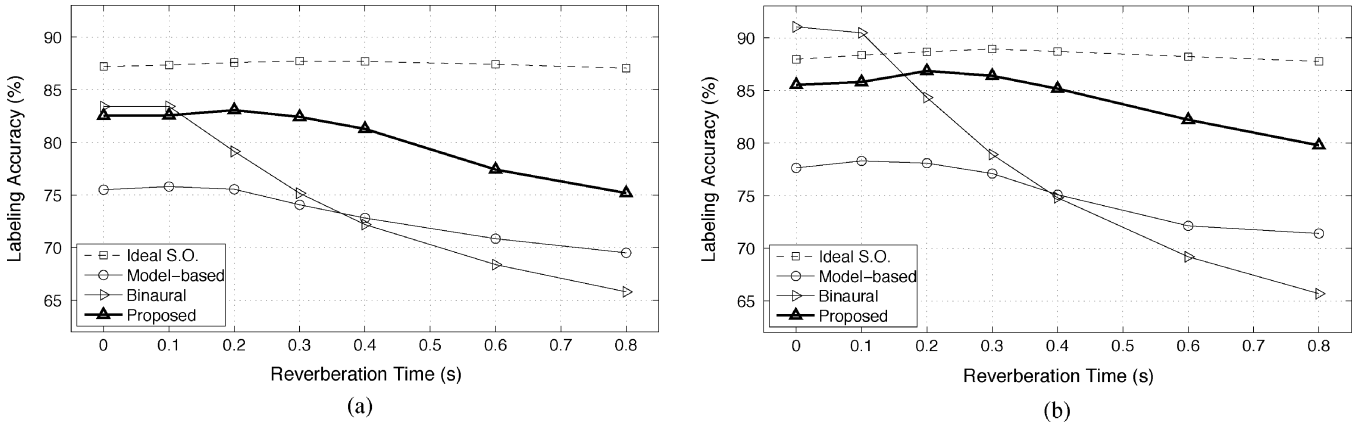


Fig. 5. Labeling accuracy of the proposed and comparison systems shown as a function of reverberation time for (a) two-talker and (b) three-talker mixtures.

TABLE I
LABELING ACCURACY AS A FUNCTION OF SPATIAL SEPARATION (IN °)

	Two-Talker Mixtures			Three-Talker Mixtures		
	≤ 30	$35-60$	>60	≤ 30	$35-60$	>60
Binaural	63.3%	74.8%	79.8%	66.8%	73.1%	79.0%
Proposed	79.9%	85.1%	85.9%	77.5%	81.1%	82.8%

mixtures between 300 ms and 800 ms T_{60}. This suggests that using monaural simultaneous organization developed specifically for reverberant environments [19] may improve performance using the proposed framework.

One can see a strong influence of the reverberation time on the binaural system. For the two-talker mixtures in which there is little reverberation present, i.e., with T_{60} of 0 and 100 ms, the binaural system outperforms even the Ideal S.O. system. This suggests that in these cases the binaural cues are more powerful than pitch-related cues for achieving simultaneous organization. However in the three-talker case and in even moderate amounts of reverberation, simultaneous organization achieved by monaural processing improves performance over exclusively binaural grouping. The gap between the Ideal S.O. system and the binaural systems increases with both the amount of reverberation and the number of talkers, indicating that the potential gain of integrating monaural and binaural processing is greater as the mixture complexity increases.

It is clear from Fig. 5 that the proposed system represents a significant improvement over the binaural system, and that the margin between the two increases as a function of T_{60}. The performance margin is also dependent on spatial separation between sources. Table I shows the average labeling accuracy of the proposed and binaural system as a function of spatial separation between the target source and the closest interference source for mixtures with T_{60} between 300 ms and 800 ms. One can see that our system's performance does not degrade as severely as the binaural system for closely spaced sources.

Due to the nature of the monaural processing used in this study, there is some influence of source gender on performance of the proposed system. For the two-talker mixtures with T_{60} between 300 ms and 800 ms, the average labeling accuracy is 81.7% for mixtures where talkers have the same gender and 85.3% when talkers have different genders. This effect is even more pronounced for the model-based system where average accuracy is 80.2% when talkers have different genders and only 68.2% for same-gender mixtures. In our two-talker database, 46% of the mixtures have sources with different genders. The difference in performance between the proposed system and comparison systems is similar for male–male and female–female mixtures.

VII. CONCLUDING REMARKS

The results in the previous section illustrate that integration of monaural and binaural analysis allows for robust localization performance, which enables sequential organization of speech in environments with considerable reverberation. The localization-based sequential organization outperforms model-based sequential organization that utilizes only monaural cues, and the proposed integration of monaural and binaural analysis outperforms an exclusively binaural approach in terms of voiced speech segregation on two- and three-talker reverberant mixtures. We have also shown that, in addition to improving segregation performance, incorporation of monaural grouping improves localization performance over three exclusively binaural methods.

The discrete azimuth space used in this study avoids two potential issues. First, the azimuth-dependent ITD-ILD likelihood functions are manageable in number (37 for each frequency channel in this study). Second, the joint search over all possible azimuths is computationally feasible. In the case of a more finely sampled or continuous azimuth space, or a localization space that includes elevation, one would need to carefully consider how to overcome both issues. To overcome the need for training an unwieldy amount of likelihood functions in a variety of acoustical conditions, parametric likelihood functions could be used without considerable performance sacrifice. In analyzing the trained ITD-ILD likelihood functions, clear patterns emerge that could be utilized to formulate a parametric model. Certain key parameters, such as the primary peak locations and spread of the distributions, could be learned from training data from a discrete set of source positions and extrapolated to a continuous space. The second issue of joint search over all possible angles in a finely sampled or continuous space could be avoided by doing an initial search in a discretized space (such as the one used here), then refining the source positions in a limited range.

The development in Section V makes two assumptions that should be carefully examined in future work. First, we propose a maximum-likelihood framework in which all sequential organizations are equally likely. For mixtures in which the input SNR is significantly different from 0 dB, maximum *a posteriori* estimation is more appropriate and it should not be assumed that $P(y)$ is uniform. Second, we assume that all simultaneous streams are conditionally independent. While this may be reasonable for simultaneous streams that are separated in time, this assumption is questionable when two simultaneous streams overlap in time. In the majority of cases, simultaneous streams that overlap in time are due to different sources. Incorporating dependence between simultaneous stream labels should improve performance, but with increased computational cost.

Finally, since the proposed system only processes voiced speech, it is essential to develop methods to handle unvoiced speech. Binaural cues are likely a powerful tool for handling unvoiced speech, which is challenging with only monaural cues (see [20]). Future work must also analyze performance with different types of interfering signals, e.g., speech babble or non-speech intrusions.

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. Pedersen for providing feedback on a preliminary draft of this manuscript.

REFERENCES

[1] *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds.. New York: Springer, 2001.

[2] *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, D. L. Wang and G. J. Brown, Eds.. Hoboken, NJ: Wiley/IEEE Press, 2006.

[3] G. J. Brown and K. J. Palomaki, "Reverberation," in *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, D. L. Wang and G. J. Brown, Eds. New York: Wiley/IEEE Press, 2006, pp. 209–250.

[4] S. Harding, J. Barker, and G. J. Brown, "Mask estimation for missing data speech recognition based on statistics of binaural interaction," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 14, no. 1, pp. 58–67, Jan. 2006.

[5] N. Roman, S. Srinivasan, and D. L. Wang, "Binaural segregation in multisource reverberant environments," *J. Acoust. Soc. Amer.*, vol. 120, no. 6, pp. 4040–4051, 2006.

[6] Y. Izumi, N. Ono, and S. Sagayama, "Sparseness-based 2CH BSS using the EM algorithm in reverberant environment," in *Proc. WASPAA*, Oct. 2007, pp. 147–150.

[7] M. I. Mandel and D. P. W. Ellis, "EM localization and separation using interaural level and phase cues," in *Proc. WASPAA*, Oct. 2007, pp. 275–278.

[8] H. Sawada, S. Araki, and S. Makino, "A two-state frequency-domain blind source separation method for underdetermined convolutive mixtures," in *Proc. WASPAA*, Oct. 2007, pp. 139–142.

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

[10] J. F. Culling and Q. S. Summerfield, "Perceptual separation of concurrent speech sounds: Absence of across-frequency grouping by common interaural delay," *J. Acoust. Soc. Amer.*, vol. 98, pp. 785–797, 1995.

[11] C. J. Darwin and R. W. Hukin, "Auditory objects of attention: The role of interaural time differences," *J. Exp. Psychol. Hum. Percept. Perform.*, vol. 25, pp. 617–629, 1999.

[12] W. M. Hartmann, "How we localize sounds," *Phys. Today*, pp. 24–29, Nov. 1999.

[13] C. J. Darwin, "Spatial hearing and perceiving sources," in *Auditory Perception of Sound Sources*, W. A. Yost, A. N. Popper, and R. R. Fay, Eds. New York: Springer, 2007, pp. 215–232.

[14] A. Shamsoddini and P. N. Denbigh, "A sound segregation algorithm for reverberant conditions," *Speech Commun.*, vol. 33, pp. 179–196, 2001.

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

[16] J. Woodruff and D. L. Wang, "On the role of localization cues in binaural segregation of reverberant speech," in *Proc. ICASSP*, Apr. 2009, pp. 2205–2208.

[17] 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.

[18] S. Vishnubhotla and C. Y. Epsy-Wilson, "An algorithm for speech segregation of co-channel speech," in *Proc. ICASSP*, Apr. 2009, pp. 109–112.

[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, pp. 625–638, 2009.

[20] G. Hu and D. L. Wang, "Segregation of unvoiced speech from nonspeech interference," *J. Acoust. Soc. Amer.*, vol. 124, pp. 1306–1319, 2008.

[21] Y. Shao and D. L. Wang, "Sequential organization of speech in computational auditory scene analysis," *Speech Commun.*, vol. 51, pp. 657–667, 2009.

[22] D. R. Campbell, The ROOMSIM User Guide (v3.3) 2004.

[23] W. G. Gardner and K. D. Martin, "HRTF measurements of a KEMAR," *J. Acoust. Soc. Amer.*, vol. 97, pp. 3907–3908, 1995.

[24] 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.

[25] 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.

[26] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," Cambridge, U.K., Tech. Rep., MRC Applied Psychology Unit, 1988.

[27] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.*, vol. 47, pp. 103–138, 1990.

[28] G. Hu and D. L. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Trans. Audio, Speech, Lang. Process.*, 2010, to be published.

[29] G. Hu, "Monaural speech organization and segregation," Ph.D. dissertation, The Ohio State Univ., Columbus, OH, 2006.

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

[31] B. Silverman, *Density Estimation for Statistics and Data Analysis*. London, U.K.: Chapman & Hall, 1986.

[32] 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, 2003.

[33] 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.

[34] 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, 2004.

[35] K. W. Wilson and T. Darrell, "Learning a precedence effect-like weighting function for the generalized cross-correlation framework," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 2156–2164, Nov. 2006.

[36] Y. Shao and D. L. Wang, "Model-based sequential organization in cochannel speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 289–298, Jan. 2006.

[37] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds. New York: Springer, 2001, ch. 8, pp. 157–180.

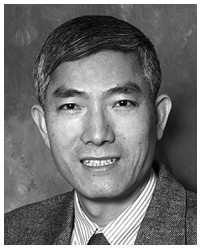
[38] C. Liu, B. C. Wheeler, W. D. O'Brien, R. C. Bilger, C. R. Lansing, and A. S. Feng, "Localization of multiple sound sources with two microphones," *J. Acoust. Soc. Amer.*, vol. 108, no. 4, pp. 1888–1905, 2000.

[39] D. L. Wang, "On ideal binary masks as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. Boston, MA: Kluwer, 2005, pp. 181–197.

[40] D. Brungart, P. S. Chang, B. D. Simpson, and D. L. Wang, "Isolating the energetic component of speech-on-speech masking with an ideal binary time–frequency mask," *J. Acoust. Soc. Amer.*, vol. 120, pp. 4007–4018, 2006.

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 (M'90–SM'01–F'04) received the B.S. and M.S. degrees from Peking (Beijing) University, Beijing, China, in 1983 and 1986, respectively, and the Ph.D. degree from the University of Southern California, Los Angeles, in 1991, all in computer science.

From July 1986 to December 1987, he was with the Institute of Computing Technology, Academia Sinica, Beijing. Since 1991, he has been with the Department of Computer Science Engineering and the Center for Cognitive Science, The Ohio State University, Columbus, where he is currently a Professor. From October 1998 to September 1999, he was a Visiting Scholar in the Department of Psychology, Harvard University, Cambridge, MA. From October 2006 to June 2007, he was a Visiting Scholar at Oticon A/S, Denmark. His research interests include machine perception and neurodynamics.

Dr. Wang received the National Science Foundation Research Initiation Award in 1992, the Office of Naval Research Young Investigator Award in 1996, and the Helmholtz Award from the International Neural Network Society in 2008. He also received the 2005 Outstanding Paper Award from the IEEE TRANSACTIONS ON NEURAL NETWORKS.