Binaural Detection, Localization, and Segregation in Reverberant Environments Based on Joint Pitch and Azimuth Cues

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

***Abstract*—We propose an approach to binaural detection, localization and segregation of speech based on pitch and azimuth cues. We formulate the problem as a search through a multisource state space across time, where each multisource state encodes the number of active sources, and the azimuth and pitch of each active source. A set of multilayer perceptrons are trained to assign time-frequency units to one of the active sources in each multisource state based jointly on observed pitch and azimuth cues. We develop a novel hidden Markov model framework to estimate the most probable path through the multisource state space. An estimated state path encodes a solution to the detection, localization, pitch estimation and simultaneous organization problems. Segregation is then achieved with an azimuth-based sequential organization stage. We demonstrate that the proposed framework improves segregation relative to several two-microphone comparison systems that are based solely on azimuth cues. Performance gains are consistent across a variety of reverberant conditions.**

***Index Terms*—Binaural speech segregation, computational auditory scene analysis, multipitch tracking, sound localization, source detection.**

I. Introduction

BINAURAL segregation and localization are important problems within computational auditory scene analysis (CASA) [35] and signal processing due to both an interest in simulating auditory perception and potential applications in hearing prostheses, robust speech recognition, spatial sound reproduction and mobile robotics. Approaches to binaural segregation assume that sound sources are separated in space and derive spatial filters to enhance the target source [3], [9], thus localization is an important subproblem. The effectiveness of spatial filtering is greatly affected by the number of microphones, the acoustic environment and the spatial configuration of sources. With only two microphones available in the binaural case, performance of segregation based on spatial cues degrades substantially as the level of reverberation increases or the separation between sources decreases [23].

In contrast, human listeners utilize both monaural and binaural cues in the perceptual organization of an acoustic scene [4]. It is not uncommon to encounter acoustic conditions with substantial reverberation, far-field sources, diffuse background noise or even co-located sources. While listeners can still achieve segregation in such cases, existing computational methods fail or experience substantial performance degradation. In order to achieve a more robust solution to the binaural segregation problem, we believe that both monaural and binaural cues should be utilized. Ideally, segregation based on monaural processing could be achieved in the absence of useful spatial cues, and the system could benefit from spatial information when available. With this idea in mind, our prior work demonstrates that both localization and segregation of voiced speech can be improved by incorporating pitch cues [38], [39]. We have shown that pitch cues can be substantially more reliable than spatial cues for across-frequency grouping (simultaneous organization), but that azimuth cues can be used reliably for grouping across time (sequential organization). These observations are consistent with aspects of auditory grouping by human listeners (see [7] for a review).

In this work we propose a binaural system for joint localization and segregation of an unknown and time-varying number of sources. We develop a novel hidden Markov model (HMM) framework to estimate the number of active sources across time, compute azimuth for each active source per frame, determine whether sources are voiced and extract pitches in voiced frames, and generate binary time-frequency (T-F) masks for the active sources. We focus on segregation of sources in fixed spatial positions, however the framework is amenable to moving sources. Whereas in our previous work we performed simultaneous organization using monaural cues and sequential organization using binaural cues in a two-stage process [38], [39], pitch and azimuth cues are considered jointly for simultaneous organization by the system proposed here. This approach retains the benefit of pitch-based grouping, but allows for improved performance when sources have similar pitches. Further, by training models jointly on pitch and azimuth cues, the relative contribution of each type of cue is learned and the system naturally deals with both voiced and unvoiced speech. This approach is motivated in part by the observation that for human listeners, monaural cues are stronger than spatial cues for simultaneous organization, but spatial cues contribute when circumstances allow (see [7], [31] for reviews).

In the following section we discuss relevant literature on binaural segregation and multichannel speech enhancement, and

Manuscript received April 17, 2012; revised July 19, 2012; accepted November 25, 2012. Date of publication December 24, 2012; date of current version January 18, 2013. This work was supported by an AFOSR grant (FA9550-08-1-0155) and a grant from the Oticon Foundation. The associate editor coordinating the review of this manuscript and approving it for publication was]Prof. Søren Holdt Jensen.

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 Center for Cognitive Science, The Ohio State University, Columbus, OH, 43210-1277 USA (e-mail: dwang@cse.ohio-state.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2012.2236316

existing work that explores processing based on both monaural and spatial cues. In Section III we describe the front-end processing, define the computational goal of the proposed system, describe the acoustic features used and provide an overview of the proposed framework. We introduce each component of the HMM framework in Section IV and describe details of the implementation used to generate experimental results in Section V. Finally, we outline the evaluation methodology and results in Sections VI and VII, and conclude with a discussion in Section VIII.

II. Background

Most binaural CASA systems use interaural (between ear) cues to perform *localization-based grouping* (see e.g., [9], [21], [28]). While numerous differences exist between localization-based grouping systems, they follow a common approach at a high level. First, left and right mixture signals are transformed into the T-F domain. Interaural cues are then extracted from each pair of T-F units. Source locations are estimated by integrating these cues across time and frequency. Once source locations are identified, predetermined models of interaural cues for the estimated source locations are used to identify the mixture T-F units that are consistent with the target location. Variants of the localization-based grouping approach that avoid prior training with a given microphone setup have also been proposed (see e.g., [23], [26]).

While the goal of many binaural segregation systems is to estimate a T-F mask, considerable effort has gone toward alternative enhancement techniques. The most ubiquitous approach to array-based enhancement is beamforming, which filters and sums the received signals in order to create a spatially-dependent attenuation pattern [3]. In principle it is possible to derive a so-called distortionless beamformer that achieves interference attenuation without degradation in the direction of interest [10]. This is in contrast to the T-F masking approach, where distortion of the target signal is unavoidable whenever attenuation is applied to a T-F unit that contains some target energy. It is possible to further increase signal-to-noise ratio (SNR) by applying a post-filter to the output of a beamformer [32]. The cascade of a distortionless beamformer and a single-channel post-filter has been shown to be mean-square error (MSE) optimal under various assumptions regarding speech distributions [14], [33]. Commonly such spatial filters are adapted across time based on target activity detection, however, optimization criteria based on higher-order statistics can also be used when the number of sources is known [5].

In spite of recent advances to more effectively deal with reverberation, underdetermined mixtures or a binaural setup, all of these approaches are inherently limited by the requirement that sound sources have sufficiently distinct spatial attributes. This requirement is not always met in practice. To address this limitation, some existing work also considers segregation based on multiple acoustic cues. Several studies have considered joint estimation of pitch and azimuth (or time delay) for one or more sources (see [16], [19], [43] as recent examples). Segregation of two talkers based on joint estimation of pitch and location using a recurrent timing neural network was proposed in [41]. The system proposed in [40] derives separate target speech estimators based on both pitch and localization cues, where estimates are then combined based on confidence scores derived from consistency of the pitch and azimuth estimates across time. Tracking of the time delay and pitch of the dominant source is handled implicitly by the system. In both [24] and [30], localization cues are used to group the harmonics of different sources across frequency, allowing for improved pitch estimation and sequential grouping of pitch points. Related approaches that incorporate binaural cues and monaural spectral models have also been proposed (see e.g., [22], [25], [37]). While many of these multi-cue approaches are relevant, we are not aware of existing methods that perform localization, pitch tracking and segregation of an unknown and time-varying number of sources.

III. Overview

A. Auditory Periphery

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 [27] with center frequencies from 80 to 5000 Hz spaced on the equivalent rectangular bandwidth scale. While the passband of the filterbank does not extend to the Nyquist frequency of 22.05 kHz, the high sample rate used facilitates computation of interaural delays with good resolution. Each bandpass filtered signal is divided into 20 ms time frames with a frame shift of 10 ms to create a cochleagram [35] of T-F units. A T-F unit is an elemental signal from one frame, indexed by m, and one filter channel, indexed by c. We use $u^E_{c,m}$ to denote the signal contained in each T-F unit, where $E \in \{L, R\}$ indicates the left or right ear signal.

B. Computational Goal

The goal of the proposed system is to segregate a desired speech signal from a binaural mixture. To perform segregation we seek to estimate the ideal binary mask (IBM) [34], which has been shown to substantially improve speech intelligibility for normal hearing and hearing-impaired listeners (see e.g., [36]). Recently, the IBM definition has been extended to deal with reverberant signals by including early reflections of the desired signal as part of the target component [29].

More formally, we model each T-F unit as,

$$u^E_{c,m} = \sum_{k=1}^{K} x^E_{k,c,m} + v^E_{c,m}, \tag{1}$$

where $x^E_{k,c,m}$ contains both the direct-path and early reflections of source k received by microphone E, $v^E_{c,m}$ denotes the combination of late reflections from all sources and any additional background noise, and K is the number of sources. Given this signal model, the so-called useful-to-detrimental ratio (UDR) [2] for source k in T-F unit $u^E_{c,m}$ can be defined as,

$$\mathrm{UDR}^E_k(c,m) = 10\log_{10}\left(\frac{\|x^E_{k,c,m}\|^2}{\|u^E_{c,m} - x^E_{k,c,m}\|^2}\right), \tag{2}$$

where $\|x\|^2$ denotes Euclidean norm, or, the energy of signal x. We then let $\mathrm{UDR}_k(c,m) = (\mathrm{UDR}^L_k(c,m) + \mathrm{UDR}^R_k(c,m))/2$ and define the IBM for source k as,

$$\mathrm{IBM}_k(c,m) = \begin{cases} 1, & \text{if } \mathrm{UDR}_k(c,m) > \mathrm{LC} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where we set LC (local SNR criterion) to 0 dB and consider the first 50 ms of the impulse response from source k to microphone E as the "useful" speech components in order to derive $x^E_{k,c,m}$[29]. Note that we average the UDR from the left and right signals so that each pair of T-F units, $u^L_{c,m}$ and $u^R_{c,m}$, are given the same assignment by $\mathrm{IBM}_k(c,m)$.

C. Feature Extraction

From each T-F unit pair, $u^L_{c,m}$ and $u^R_{c,m}$, we extract a set of pitch- and azimuth-related features as observations within the tracking framework described in Section IV. The pitch-related features are based on the correlogram and envelope correlogram [35]. The correlogram, denoted $A^E(c,m,\gamma)$, is a normalized running auto-correlation performed in individual frequency channels for each time frame. The envelope correlogram, denoted $\bar{A}^E(c,m,\gamma)$, is the same, but envelope extraction is performed prior to computation of the auto-correlation. Also note that we decimate the left and right signals to 16 kHz before computation of the correlograms. We then let

$$\chi_{c,m}(\gamma) = \{A^L(c,m,\gamma), A^R(c,m,\gamma), \bar{A}^L(c,m,\gamma), \bar{A}^R(c,m,\gamma)\}$$

denote the set of four pitch-related features for channel c, frame m and lag γ. We use X_m to denote the full set of pitch features for frame m.

The binaural features calculated are the interaural time difference (ITD), denoted $\tau_{c,m}$, and the interaural level difference (ILD), denoted $\lambda_{c,m}$. We calculate ITD as the maximum peak in a running cross-correlation between $u^L_{c,m}$ 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 [38]. We use T_m and Λ_m to denote the full set of ITD and ILD features, respectively, for frame m. Finally, we use $Z_m = \{T_m, \Lambda_m, X_m\}$ to denote the entire set of observed data for frame m.

D. Overview of the Proposed Framework

We utilize both spatial and periodicity information to estimate $\mathrm{IBM}_k(c,m)$. To do so we track the pitch and azimuth of multiple concurrent sources across time. We formulate the tracking problem such that we attempt to identify the most probable path through a multisource state space, where a multisource state encodes the number of active sources, the azimuth of each active source, and the voicing of each active source. By identifying a path through the multisource state space, we then generate a solution to the detection, localization and pitch estimation problems. Further, we incorporate a set of multilayer perceptrons (MLPs) trained jointly on pitch and azimuth cues in order to assign T-F units to one of the active sources in each multisource state. Essentially, the MLPs generate frame-level T-F masks, which are then stitched together across time when the multisource state path is determined. In this way, the system also generates a solution to the simultaneous organization problem (across-frequency grouping within a continuous time interval). Finally, azimuth-based sequential organization is performed to generate a T-F mask for each detected source.

As will be discussed in Section V-A, the cardinality of the full multisource state space is prohibitively large. In order to make computation feasible, we incorporate independent pitch and azimuth modules to identify a set of pitch and azimuth candidates to be considered by the HMM in each frame. We first introduce the main components of the HMM in Section IV, and then describe how the independent modules are used to generate candidate states in Section V-A.

IV. Hidden Markov Model Framework

We seek to model the posterior probability of a multisource state in each time frame based on the observed features described in Section III-C. A multisource state, denoted $S = \{\theta_1, \ldots, \theta_K, \gamma_1, \ldots, \gamma_K\}$, is a collection of individual pitch and azimuth states for K sources. We consider a discrete grid of azimuths from -90° to 90° and allow sources to be inactive, where we let $\theta = \emptyset$ denote an inactive source. Similarly, we consider a discrete set of pitch lags and allow sources to be unvoiced, where we let $\gamma = \emptyset$ denote an unvoiced source.

The posterior probability of a multisource state given the observed data can be expressed as,

$$p(S_m|Z_{1:m}) \propto p(Z_m|S_m)p(S_m|Z_{1:m-1}) \tag{4}$$

where subscript $_{1:m}$ denotes a collection of features from frame 1 through frame m. In the subsections below we discuss computation of the observation likelihood, $p(Z_m|S_m)$, and the state predictor, $p(S_m|Z_{1:m-1})$. We first discuss a data association stage where, in keeping with the assumption made by the IBM that each T-F unit can be assigned to at most one source, T-F units are assigned to an individual source for each hypothesized multisource state. These assignments are the mechanism by which T-F masks are generated and also facilitate computation of the observation likelihood.

A. T-F Unit Assignment

As stated above, for each possible multisource state in each time frame, we assign T-F units to one of the active sources. Once the state path is determined the T-F unit assignments associated with the selected multisource states are used to construct binary T-F masks. One of the principal advantages of the binary T-F masking approach to speech segregation is that it opens up a class of supervised learning algorithms to perform classification. Recent methods have yielded promising results using binaural features [28], pitch-based features [17] and amplitude modulation features [20]. In keeping with this work, we incorporate a set of trained MLPs to assign T-F units to individual sources based on the azimuth and pitch information contained in a multisource state.

Specifically, let $H_{k,c,m}$ denote the hypothesis that source k, with azimuth θ_k and pitch γ_k, satisfies the criteria necessary to be labeled 1 by the IBM (see (3)). We then let $p_c(H_{k,c,m}|z_{c,m},\theta_k,\gamma_k)$ denote the posterior probability of $H_{k,c,m}$ given the monaural and binaural features, $z_{c,m}$. For a given multisource state, we perform data association according to,

$$y_{c,m}(S_m) = \underset{k\in\{1,\ldots,K\}}{\arg\max}\ [p_c(H_{k,c,m}|z_{c,m},\theta_k,\gamma_k)]\,. \tag{5}$$

We train a set of MLPs to model $p_c(H_{k,c,m}|z_{c,m},\theta_k,\gamma_k)$. The models are frequency-, azimuth- and pitch-dependent. However, since the correlogram features, $\chi_{c,m}(\gamma)$, are a function of

pitch lag we do not train separate models for each possible pitch (see [17]), but rather train two MLPs for each frequency channel and azimuth. Two models per channel and azimuth are necessary to accommodate both voiced and unvoiced sources. When a source is unvoiced (i.e., $\gamma_k = \emptyset$) we let,

$$p_c(H_{k,c,m}|z_{c,m}, \theta_k, \gamma_k) = g^B_{c,\theta_k}(\tau_{c,m}, \lambda_{c,m}), \quad (6)$$

and when a source is voiced we let,

$$p_c(H_{k,c,m}|z_{c,m}, \theta_k, \gamma_k) = g^J_{c,\theta_k}(\tau_{c,m}, \lambda_{c,m}, \chi_{c,m}(\gamma_k)). \quad (7)$$

Here, $g^B_{c,\theta}(\cdot)$ denotes the output of a *Binaural MLP* (i.e., trained on only ITD and ILD features) for frequency channel c and azimuth θ, and $g^J_{c,\theta}(\cdot)$ denotes the output of a *Joint MLP* (i.e., trained on joint monaural and binaural features) for frequency channel c and azimuth θ. Details regarding the training procedures, training data and MLP topology are described in Section VI-C.

B. Observation Likelihood

The multisource observation likelihood captures the probability that the observed features were generated by a set of sources with azimuth and pitch characteristics specified by the multisource state S_m. We first assume that binaural features and pitch-related features are conditionally independent such that, $p(Z_m|S_m) = p(T_m, \Lambda_m|S_m)p(X_m|S_m)$.

To calculate $p(T_m, \Lambda_m|S_m)$ we further assume conditional independence across frequency channels so that the T-F unit assignment expressed by (5) allows for the decomposition of frame-level likelihoods conditioned on the characteristics of multiple sources into the product of unit-level likelihoods conditioned on the characteristics of a single source. Accordingly, we let,

$$p(T_m, \Lambda_m|S_m) = \alpha(S_m)\left(\prod_c p_c(\tau_{c,m}, \lambda_{c,m}|\theta_{y_{c,m}(S_m)})\right)^{\xi}, \quad (8)$$

where $\alpha(S)$ is used to adjust the likelihoods based on the number of active sources contained in S, and the term ξ is used to overcome the so-called probability overshoot phenomenon [13]. Multisource states with more active sources will produce systematically higher likelihoods due to increased flexibility in the T-F unit assignment expressed by (5). The penalty term serves to minimize any systematic bias towards overestimating the number of sources. Probability overshoot results from the fact that, due to the overlapping passbands of gammatone channels, observations in individual channels are not entirely independent.

For the multisource state with no active sources, $S_m = \{\emptyset, \ldots, \emptyset\}$, we set $p(T_m, \Lambda_m|S_m) = P_{background}$. Together, $P_{background}$ and $\alpha(S_m)$ control the detection sensitivity of the system. Finally, we use the azimuth-dependent GMMs proposed in [39] for $p_c(\tau_{c,m}, \lambda_{c,m}|\theta)$. Note that while $p(T_m, \Lambda_m|S_m)$ depends on the full multisource state (i.e., azimuths, pitches and the azimuth-pitch correspondence) due to the use of $y_{c,m}(S_m)$, the likelihood of a specific pair of ITD and ILD values, $\tau_{c,m}$ and $\lambda_{c,m}$, is assumed to be independent of a source's pitch.

TABLE I
SINGLE SOURCE STATE TRANSITION PROBABILITIES. ROWS 1, 2 AND 3 LIST TRANSITIONS OUT OF VOICED, UNVOICED AND INACTIVE STATES, RESPECTIVELY. COLUMNS 1, 2 AND 3 LIST TRANSITIONS INTO VOICED, UNVOICED AND INACTIVE STATES, RESPECTIVELY.

	$p(\theta, \gamma\|\cdot)$	$p(\theta, \emptyset\|\cdot)$	$p(\emptyset, \emptyset\|\cdot)$
$p(\cdot\|\theta', \gamma')$	$P_{\sim d}P_{vv}f(\gamma\|\gamma')f(\theta\|\theta')$	$P_{\sim d}P_{vu}f(\theta\|\theta')$	P_d
$p(\cdot\|\theta', \emptyset)$	$P_{\sim d}P_{uv}p(\gamma)f(\theta\|\theta')$	$P_{\sim d}P_{uu}f(\theta\|\theta')$	P_d
$p(\cdot\|\emptyset, \emptyset)$	$P_bP_vp(\gamma)p(\theta)$	$P_bP_{\sim v}p(\theta)$	$P_{\sim b}$

We take a relatively simple approach to calculate $p(X_m|S_m)$. As will be discussed further in Section V-A, we incorporate individual pitch and azimuth modules to supply a small set of candidate multisource states to the HMM. The likelihood of pitch combinations, $p(X_m|\Gamma_m)$, where $\Gamma_m = \{\gamma_1, \ldots, \gamma_K\}$, is computed within the pitch module. Since the joint dependence on pitch and azimuth is already captured by $p(T_m, \Lambda_m|S_m)$, we simply let $p(X_m|S_m) = p(X_m|\Gamma_m)$, which is described in Section V-A. Essentially, $p(X_m|S_m)$ captures the overall salience of a given set of pitches, independent of how they are paired with azimuths in S_m, and $p(T_m, \Lambda_m|S_m)$ then validates both the salience of an azimuth set and the pitch-azimuth correspondence specified by S_m.

C. State Predictor

The state predictor captures the probability of a given multisource state given the posterior probabilities from the previous frame and state transition probabilities. To estimate the optimal *path* through the multisource state space using the Viterbi algorithm, we approximate the predictor using,

$$p(S_m|Z_{1:m-1}) \approx \max_{S_{m-1}} \left[p(S_m|S_{m-1})p(S_{m-1}|Z_{1:m-1})\right], \quad (9)$$

and keep track of the prior state, S_{m-1}, that maximizes the right-hand side of (9) as the *predecessor* of S_m, denoted $\Pr(S_m)$. We assume independence between source azimuths and pitches and define the multisource state transition probabilities according to,

$$p(S_m|S_{m-1}) = \prod_{k=1}^{K} p(\theta_{m,k}, \gamma_{m,k}|\theta_{m-1,k}, \gamma_{m-1,k}). \quad (10)$$

We list individual state transition probabilities in Table I where P_b and P_d are birth and death probabilities, respectively, $f(\theta|\theta')$ denotes the azimuth transition probability, $f(\gamma|\gamma')$ the pitch transition probability, P_v the prior probability of a source being voiced, and P_{vv} and P_{uu} are the voiced-voiced and unvoiced-unvoiced transition probabilities, respectively. Also, we let $P_{\sim b} = 1 - P_b$, $P_{\sim d} = 1 - P_d$, $P_{\sim v} = 1 - P_v$, $P_{vu} = 1 - P_{vv}$ and $P_{uv} = 1 - P_{uu}$. P_b, P_d, $f(\theta|\theta')$ and $p(\theta)$ are highly situation dependant, as they are related to source activity, source motion and listener movements. In contrast, P_v, P_{vv}, P_{uu}, $f(\gamma|\gamma')$ and $p(\gamma)$ capture general properties of speech and should be relatively consistent across conditions. We describe how these parameters are set in Section V-C.

V. Implementation

Summary of the HMM framework

Initialize $p(\{\emptyset, \ldots, \emptyset\}) = 1$;
for $m \leftarrow 1$ **to** M **do**
Compute X_m, T_m and Λ_m; // **Section III-C**
Construct multisource candidates, $\hat{S}_m$; // **Section V-A**
foreach $S_m \in \hat{S}_m$ **do**
Compute $y_c(S_m)$; // **Section IV-A**
Compute $p(Z_m|S_m)$; // **Section IV-B**
Compute $p(S_m|Z_{1:m-1})$ and $\Pr(S_m)$; // **Section IV-C**
Compute $p(S_m|Z_{1:m}) = p(Z_m|S_m)p(S_m|Z_{1:m-1})$;
end
end
Determine $\hat{S}_{1:M}$ and perform simultaneous organization;
Perform sequential organization; // **Section V-B**

In this section we describe several implementation choices made for the experiments presented in Section VII. Two primary issues must be addressed in order to use the HMM framework described in the previous section for segregation. First, as noted in Section III, a full search through the HMM state space is intractable. The cardinality of S is $((|\Theta|-1)|\Gamma|+1)^K$, which is roughly 250 billion states for the number of sources and the pitch-azimuth grid used in our experiments. Second, the HMM performs simultaneous organization in that grouping is performed across frequency over continuous time intervals. If an active source becomes silent and then reappears at a later time, the model is agnostic as to whether the two periods of activity are due to the same source. To perform segregation it is necessary to include a subsequent sequential organization stage that links periods of source activity across disconnected time intervals.

In the following subsections we propose methods to achieve these two ends. To reduce the HMM search space we incorporate independent pitch and azimuth modules to identify a set of pitch and azimuth candidates for each frame. We then propose a simple azimuth-based sequential organization stage. It is important to keep in mind that effective solutions to these problems depend on time and memory constraints, whether or not an online solution is required, or assumptions regarding source activity and source motion. The implementation choices described in this section are suitable for the experiments conducted, but could easily be replaced as needed for alternative applications. We provide a summary of the proposed HMM framework above.

A. Pitch and Azimuth Modules

To make computation feasible we incorporate independent pitch and azimuth modules to identify a set of pitch and azimuth candidates for each frame. We use the multipitch tracking system proposed in [18] as the independent pitch module. We let $p(\Gamma_m|X_{1:m})$[1] denote the posterior probability of a multipitch state in frame m, and let $p(X_m|\Gamma_m)$ denote the multipitch likelihood. Note that the Jin-Wang multipitch tracking system explicitly deals with up to two simultaneous voiced sources. As a result, the proposed framework, while it can localize and segregate up to K sources, will assign a pitch to at most two sources. We discuss this limitation of the current implementation further in Section VIII.

The azimuth module is a version of the full HMM that ignores correlogram features, or essentially, assumes all sources are unvoiced. We compute $p(\Theta_m|T_{1:m}, \Lambda_{1:m})$, where $p(T_m, \Lambda_m|\Theta_m)$ is calculated according to (8) and T-F unit assignments are computed using (5). In this case, posterior probabilities are modeled with only the binaural MLPs (see (6)) since all sources are assumed to be unvoiced. We again assume independence between sources to calculate transition probabilities, and individual source azimuth transitions, $p(\theta_{m,k}|\theta_{m-1,k})$, are computed by marginalizing the transition probabilities expressed in Table I[2].

Once $p(\Gamma_m|X_{1:m})$ and $p(\Theta_m|T_{1:m}, \Lambda_{1:m})$ are computed for frame m, we use them to identify a set of multipitch candidates, $\hat{\Gamma}_{\mathrm{m}}$, and multiazimuth candidates, $\hat{\Theta}_{\mathrm{m}}$. A set of multisource candidates, $\hat{S}_{\mathrm{m}}$, are then created by considering all valid combinations[3] of multipitch and multiazimuth candidates.

B. Segregation

While online tracking and segregation of moving sources are possible with the proposed framework, we focus on offline segregation of an unknown number of sources in fixed spatial positions in the experiments presented in Section VII. This facilitates comparison to existing blind source separation (BSS) methods that assume a known number of spatially fixed sources and utilize the full mixture to localize and separate each source [8], [23].

We first determine the optimal path through the multisource state space using the Viterbi algorithm. A state sequence, denoted $\hat{S}_{1:M}$ where M is the total number of frames, encodes when individual sources become active and inactive, and encodes the azimuth, voicing characteristics and set of T-F units associated with each source while it is active. When moving across time through the identified state path, we begin a new simultaneous stream[4], pitch contour and azimuth contour from the frame-level T-F mask, pitch estimate and azimuth estimate associated with the new source in the multisource state. The stream and associated contours are propagated across time until the source becomes inactive.

Since we assume sources are in fixed spatial positions in this study, azimuth is a powerful cue for sequential organization. As such, subsequent to the formation of simultaneous streams we label a stream as target dominant when its associated azimuth is within a specified error tolerance around the known target azimuth.

[1]Note that we use notation consistent with Section IV rather than the notation used in [18].

[2]$p(\theta|\theta') = P_{\sim d} f(\theta|\theta')$, $p(\theta|\emptyset) = P_b p(\theta)$, $p(\emptyset|\theta) = P_d$, $p(\emptyset|\emptyset) = P_{\sim b}$

[3]The number of pitches cannot exceed the number of active sources.

[4]A group of T-F units over a continuous time interval.

C. Parameter Settings

Several parameters were introduced in the development of the HMM framework presented in Section IV. We now list the parameter settings chosen for the experiments performed and provide the motivation for our choices.

In all of the experiments performed we set $K = 3$, meaning that the system will handle up to three simultaneous sources. We use a discrete grid of azimuths in steps of 5° from $-90°$ to 90° and a discrete grid of pitch lags from 32 to 200 samples (80 to 500 Hz at 16 kHz sample rate). The azimuth grid is dictated by the methods used for binaural simulation, which are described in Section VI-A. The pitch grid is motivated by the range of common fundamental frequencies for speech signals. Consistent with our pitch search range, we use a low-pass filter with 500 Hz cutoff frequency and a Kaiser window to extract signal envelopes for correlogram computation (see Section III-C).

For computation of the observation likelihood in (8), $\alpha(S)$ and $P_{background}$ are free parameters that control source detection sensitivity. ξ is a free parameter that helps balance observation and transition probabilities. Each of these values was determined using a held out validation set. We set $\alpha(S)$ to 1, 1, 0.4, and 0.25 for the cases with 0, 1, 2 and 3 active sources contained in S, respectively, set $P_{background} = 0.02$, and set $\xi = 1/16$.

Since we consider spatially fixed sources in the evaluation, we set $f(\theta|\theta') = \delta(\theta - \theta')$. We assume a uniform prior of azimuth and set $p(\theta) = 1/(|\theta| - 1)$. Again based on the validation set, we set $P_b = 0.03$ and $P_d = 0.01$. Based on a set of utterances from the TIMIT corpus [12], we set $P_v = 0.71$, $P_{vv} = 0.97$, $P_{uu} = 0.91$. Following [18], [42] we use a Laplacian distribution with mean 0.4 and standard deviation 2.4 for $f(\gamma|\gamma')$.

Finally, in order to construct the set of multisource candidate states, $\hat{S}_m$, we choose a set of multipitch candidates, $\hat{\Gamma}_{\mathrm{m}}$, and multiazimuth candidates, $\hat{\Theta}_{\mathrm{m}}$, according to $p(\Gamma_m|X_{1:m})$ and $p(\Theta_m|T_{1:m}, \Lambda_{1:m})$, respectively. In preliminary experiments we find that the most common source of errors in multipitch tracking is correctly identifying the true number of pitches (this observation is consistent with the performance reported in [18]). With this in mind, we select the three best 1-pitch and ten best 2-pitch candidates based on $p(\Gamma_m|X_{1:m})$ and also allow for the possibility of no voiced source in each frame. This yields a total of 70 multipitch candidate states once we allow for the possibility of pitch assignment to any of the three sources. We select the top 150 multiazimuth candidates according to $p(\Theta_m|T_{1:m}, \Lambda_{1:m})$. When combined to form $\hat{S}_m$, these settings yield a total of 10,500 multisource candidate states in each frame, a reduction of over 7 orders of magnitude relative to the full multisource state space. In preliminary experiments we found the system to be relatively insensitive to the number of multipitch and multiazimuth candidates considered, and that good tracking and segregation performance could be achieved even with such a severe reduction in the search space.

VI. Evaluation Methodology

A. Binaural Simulation

For both the training and evaluation databases, we generate binaural mixtures that simulate pickup of multiple speech sources in a reverberant space. Speech signals are drawn from the TIMIT database [12] and passed through a binaural impulse response (BIR) for a specified angle and room condition. We use both simulated and measured BIRs. We simulate BIRs with the ROOMSIM package [6]. We generate BIRs with T_{60} equal to 0.2, 0.4 and 0.6 s, where for each T_{60} we create 15 room environments where room size, microphone position and microphone orientation are selected randomly and the reflection coefficients of wall surfaces are set to be equal and the same across frequency. For each T_{60} and environment we create BIRs for source positions between $-90°$ and 90°, spaced by 5°, where the source is placed 2 m from the microphone array. Anechoic HRTF measurements from a KEMAR mannequin [11] are used in the simulation, so we refer to the simulated set of BIRs as the *KEMAR BIRs*. The measured BIRs are described in [15]. 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 refer to this set as the *HATS BIRs*.

B. Evaluation Database

To evaluate the proposed system we generate three sets of mixtures that cover a variety of acoustic conditions. Since an important component of the proposed system is estimating the number of active speech sources across time, we interlace monaural utterances from the same TIMIT speaker with periods of silence to form an individual speech source. Specifically, for each source we randomly choose an initial silence period between 0.1 and 1 s, a speech duration between 1 and 2 s and a gap duration between 0.1 and 1.5 s. Given these values a source is created by first placing zeros in the signal for the initial silence, then alternating between speech and silence periods until a 3 s signal has been created. Random utterances (without duplication) from the same speaker are used for all speech periods of the same source, but TIMIT speakers are chosen at random for each mixture. This process is carried out with monaural TIMIT signals prior to spatialization and ensures that each mixture contains a time-varying number of sources.

1) Set 1: For evaluation set 1 we simulate two speech sources at four different angular separations. For all mixtures we use the KEMAR BIRs with T_{60} set to 0.4 s and place a target source at 0°. We place the interference source at 5°, 10°, 15°, or 30°. We generate 25 mixtures for each condition. The spatialized sources are set to have equal power when summated across left and right signals. To simulate a small amount of diffuse background noise, we filter uncorrelated speech-shaped noise through the anechoic BIRs for each azimuth ($-90°$ to 90°) and sum them together. We create the speech-shaped filter by averaging the amplitude spectra of 200 speech utterances drawn from TIMIT at random. We then add the diffuse noise to each mixture such that the total speech-to-noise ratio is 24 dB.

2) Set 2: For evaluation set 2 we generate both two- and three-talker mixtures where the azimuth of each source is selected randomly such that sources are spaced by 10° or more.

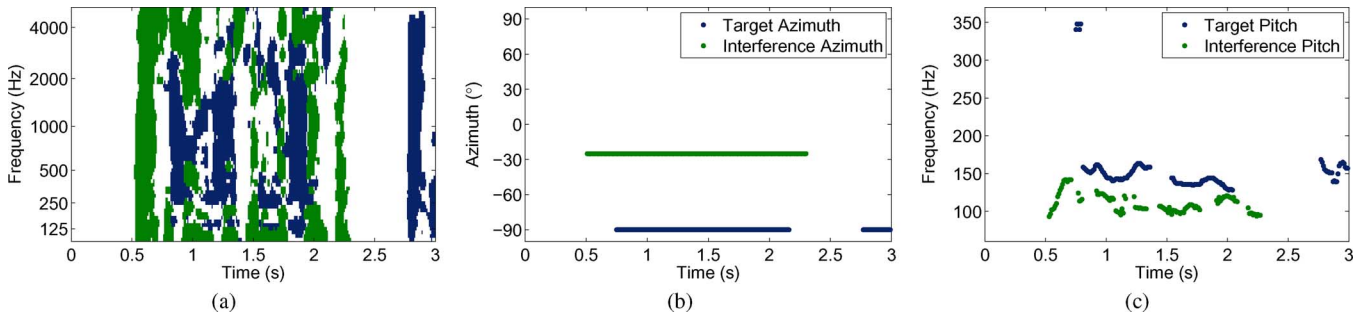

Fig. 1. Example IBMs (a), ground truth activity and azimuth (b), and ground truth pitch (c) for a mixture of two male talkers from evaluation set 2. Target mask and associated pitch and azimuth estimates are shown in blue, interference in green. (a) Ideal binary masks. (b) Ground truth source activity and azimuth. (c) Ground truth pitch.

We again use simulated BIRs in order to control T_{60}. We generate 100 mixtures for the two- and three-talker cases with T_{60} equal to 0.2, 0.4 and 0.6 s. Source distance is set to 2 m for all sources. Spatialized sources are set to have equal power and diffuse noise is added to achieve a 24-dB speech-to-noise ratio.

3) Set 3: To evaluate the system using real impulse responses we generate 50 two-talker mixtures for each room environment contained in the HATS BIR set. Azimuths are selected randomly such that sources are spaced by 10° or more. Again, spatialized sources have equal power and diffuse noise is added to achieve a 24-dB speech-to-noise ratio.

C. Model Training

The proposed system utilizes trained models in both the observation likelihood (see Section IV-B) and generation of T-F masks (see Section IV-A). We use the GMMs described in [39] to compute observation likelihoods. Two sets of GMMs are incorporated: one for evaluation sets 1 and 2 (KEMAR GMMs) and one for evaluation set 3 (HATS GMMs).

For the MLPs used in T-F mask generation, we also train separate models for evaluation sets 1 and 2 (KEMAR MLPs) and evaluation set 3 (HATS MLPs). To train the KEMAR MLPs we generate 100 mixtures for each azimuth between $-90°$ and $-90°$ where the number of interfering talkers (up to 4), interference azimuths, source distances and mixture T_{60} are selected randomly. The room environments used for the training simulations are different from those used in evaluation sets 1 and 2. To train the HATS MLPs we generate 100 mixtures for each azimuth in each of the four reverberant room conditions (rooms A, B, C and D; see Section VI-A). The number of interfering talkers (up to 4) and interference azimuths are selected randomly for each mixture. A separate set of models is trained for each room condition (e.g., room A) on the data from the three alternative rooms (e.g., rooms B, C and D) so that the impulse responses used in an evaluation utterance have not been seen in training.

For each mixture we generate the observed binaural and monaural features (see Section III-C), calculate the IBM according to (3) and extract the ground truth pitch of the target source from the premixed signals using the method proposed in [1]. The IBM provides the desired classification label and pitch information is used to select the appropriate correlogram features and to distinguish between voiced and unvoiced target frames. We train a binaural MLP for each frequency and azimuth using the ITD and ILD features. Using only voiced target frames, we train a joint MLP for each frequency and azimuth using ITD, ILD and correlogram features.

For simplicity each MLP has the same network topology consisting of a hidden layer with 20 nodes, and hyperbolic tangent sigmoid transfer functions for both hidden and output nodes. Training is accomplished using a generalized Levenberg-Marquardt backpropagation algorithm.

VII. Evaluation

A. Experiment 1: Comparison With Ground Truth Information

In this experiment we validate the fundamental assumption that segregation based jointly on pitch and azimuth outperforms segregation based on azimuth alone. To do so we compare the quality of binary T-F masks generated using the MLPs that consider both correlogram and binaural features versus the MLPs that rely only on ITD and ILD. We measure performance using the difference between the percentage of correctly labeled target-dominant units (Hit) and the percentage of incorrectly labeled interference-dominant units (FA), or Hit-FA, which has been shown to correlate well with speech intelligibility [20]. We show Hit-FA results assuming ground-truth pitch and azimuth to establish the ceiling performance achievable by the proposed mask estimation methods. We also show results with estimated pitch and azimuth to analyze the amount of degradation due to estimating the number of sources and the corresponding pitches and azimuths across time.

We perform this set of experiments using evaluation set 1. For each mixture we generate the IBM for each source according to (3) and use the pitch tracking approach proposed in [1] on the premixed signals to generate ground truth pitch for each source. The IBM, ground truth pitch and known azimuth allow us to generate ground truth frame-level labels for each source. We consider a source to be active in a frame when at least one T-F unit in the source's IBM is labeled 1. Each active frame is labeled with the source's known azimuth to generate ground truth azimuth for each source. For each active frame, we label the frame as either voiced or unvoiced depending on whether a pitch has been detected. The ground truth pitch for each source is then associated with that source's voiced frames. We show the IBMs, ground truth source activity and azimuth, and ground truth pitch points for an example mixture with two male talkers in Fig. 1.

In Table II we show the average Hit-FA for the proposed system ("Estimated Azimuth+Pitch") along with the system

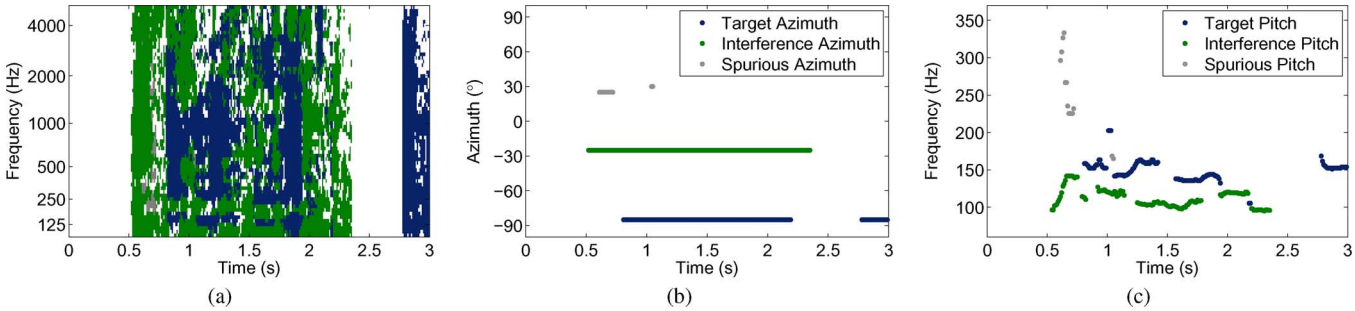

Fig. 2. Example estimated masks (a), source activity and azimuth (b), and pitch (c) for a mixture of two male talkers from evaluation set 2. Target mask and associated pitch and azimuth estimates are shown in blue, interference in green. Estimates due to a spurious (falsely detected) source are shown in gray. (a) Estimated binary masks. (b) Estimated source activity and azimuth. (c) Estimated pitch.

TABLE II
AVERAGE HIT-FA (%) ON EVALUATION SET 1 FOR THE PROPOSED AND THE AZIMUTH-ONLY SYSTEM WITH GROUND-TRUTH (G.T.) AND ESTIMATED PITCH/AZIMUTH. TARGET AT 0° FOR ALL MIXTURES.

	Interference Azimuth				
	5°	10°	15°	30°	Avg.
G.T. Azimuth+Pitch	74.4	75.0	75.9	77.5	75.7
Estimated Azimuth+Pitch	67.3	72.4	73.5	75.2	72.1
G.T. Azimuth	59.4	63.6	65.7	69.6	64.6
Estimated Azimuth	54.9	61.6	63.8	67.5	61.9

that incorporates ground-truth azimuth and pitch ("G.T. Azimuth+Pitch"). We also show performance using azimuth cues alone, based on ground truth ("G.T. Azimuth") and estimated ("Estimated Azimuth") azimuth. The azimuth-only version is achieved by the azimuth module described in Section V-A. The ceiling performance achievable by the proposed MLPs is shown in the first row. We see that the decrease in Hit-FA as the interference source is placed more closely to the target source is roughly 3%. Hit-FA is systematically lower for the ground truth azimuth-only system (11.1% drop on average from row 1 to row 3), and the degradation between the 5° case and 30° case is roughly 10%. Consistent with the ground-truth case, the proposed system based on both pitch and azimuth achieves a systematic improvement relative to the azimuth-only system. We see an improvement of up to 12.4% in the 5° case and 10.2% on average.

We show the output of the proposed system on an example mixture from evaluation set 2 in Fig. 2. Comparing Fig. 1 to Fig. 2 illustrates the high degree of accuracy in mask estimation and pitch/azimuth estimation. The majority of errors in mask estimation are due to falsely detected T-F units in regions dominated by reverberation (and thus labeled 0 in the IBMs). Reverberation tends to smear the periodic speech components across time and thus some T-F units in the reverberation tail are incorrectly assigned to the detected sources.

B. Experiment 2: Comparison to Existing Systems

In this experiment we compare the proposed approach to three two-microphone systems from the literature. The first is a fixed MVDR beamformer [3]. To ensure comparison to a beamformer with good performance, we calculate 256-tap filters using the clean target and residual signals. We decimate the mixture to 16 kHz prior to filtering. We also compare our method to the recent segregation methods presented in [8], [23]. Both of these methods assume the number of sources is known and that sources are in fixed spatial locations. Although not required by the proposed approach, we provide these comparison methods with the number of speech signals contained in each mixture. We note that the method proposed in [8] was not explicitly designed to handle binaural mixtures, and thus is sensitive to spatial aliasing caused by a large microphone spacing. This approach is representative of a class of BSS methods that handle underdetermined mixtures by performing separation independently in each frequency band and then seek to resolve the across-frequency permutation ambiguity. We include these results to illustrate the difficulty the binaural case poses to solving the permutation problem.

In Fig. 3, we show the change in SNR (ΔSNR) achieved by the proposed and comparison systems on evaluation sets 1 and 2. Since the comparison systems do not estimate the IBM, we use the premixed target signal as the reference in calculation of SNR. Note that, in keeping with our definition of the IBM, we include early reflections as a part of the target signal. In Fig. 3(a) we see that the proposed approach achieves an improvement in terms of ΔSNR relative to the comparison methods for all angular separations between target and interference. The improvement is largest for the 5° and 10° mixtures, where it exceeds 3 dB. In Fig. 3(b) and (c) we show ΔSNR achieved on evaluation set 2 as a function of T_{60} for two- and three-talker mixtures, respectively. The proposed system achieves the largest SNR gains in nearly all cases.

As one would expect, the MVDR is able to achieve much larger SNR gains for mixtures with two talkers, particularly when there is little reverberation, because it is able to create a single null in the interference direction. As reverberation increases, sources are spaced more closely or the number of talkers is increased, the beamformer is less effective. Using longer filters could allow for improved interference rejection, but in preliminary experiments we have also found an MVDR estimated from the mixture signal based on target detection is considerably less effective. However, since performance is influenced by numerous factors such as the activity detection method, the filter length and the amount of averaging used to derive the beam pattern, we include only the MVDR results based on premixed signals in this comparison.

The Duong *et al.* system is an iterative implementation of the multichannel Wiener filter that combines a beamformer and

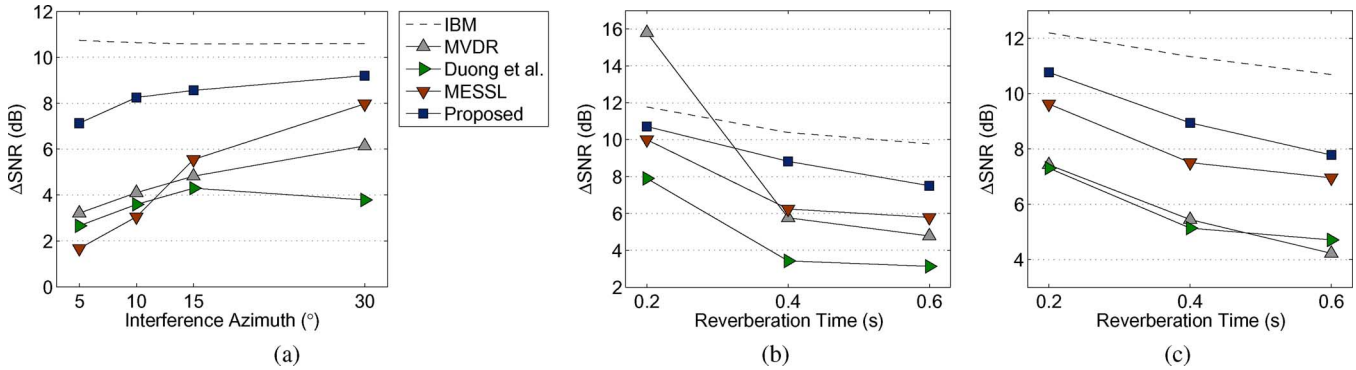

Fig. 3. Avg. ΔSNR for the proposed algorithm and three comparison methods on evaluation sets 1 (a) and 2 (b,c).

TABLE III
AVG. ΔSNR (IN dB) FOR THE PROPOSED AND THREE COMPARISON SYSTEMS USING MEASURED IMPULSE RESPONSES FROM FOUR ROOM CONDITIONS. T_{60} FOR EACH ROOM (IN s) IS LISTED IN PARENTHESIS.

	A (0.32)	B (0.47)	C (0.68)	D (0.89)	Avg.
Proposed	8.8	7.6	8.9	6.9	8.1
MVDR	5.5	4.4	5.4	4.1	4.9
Duong *et al.*	3.5	3.1	3.9	3.5	3.5
MESSL	5.8	5.3	6.6	6.4	6.0
IBM	11.3	10.0	10.8	9.6	10.4

post-filter. This system does not perform well on our evaluation set due to the large distance between microphones. As mentioned above, it is important to note that the authors did not design the system for such a large microphone spacing and thus our result should not be too surprising, but it does illustrate the challenge in resolving the across-frequency permutation ambiguity for a binaural input.

The MESSL system outperforms the other comparison methods and is capable of achieving large gains in SNR when sources are well separated in space. This is notable particularly because MESSL requires little prior training and is still capable of handling spatial aliasing.

In Table III we show ΔSNR achieved by the proposed and comparison systems on evaluation set 3, which uses measured BIRs in four different room conditions. Consistent with the results when using simulated BIRs, the proposed system achieves the best performance in all room conditions.

VIII. CONCLUDING REMARKS

The evaluation results show that the proposed integration of pitch and azimuth cues achieves more robust segregation than considering azimuth cues in isolation. Experiment 1 makes a direct comparison between the binary masks estimated by the system with and without pitch cues and shows a marked improvement in estimated masks through inclusion of pitch. Experiment 2 shows that the proposed method outperforms three existing two-microphone systems from the literature. Improvement relative to the comparison methods of [8], [23] is notable given that these methods assume that the number of sources is known, while sources are detected by the proposed method. Further, while [8], [23] assume that sources are fixed, the proposed method is capable of processing mixtures with moving sound sources by altering $f(\theta|\theta')$.

One shortcoming of the proposed approach is that only two simultaneous pitches can be tracked. We extended the pitch module to handle three concurrent pitches but found little performance gain. Analysis revealed that computing observation likelihoods as described in [18] did not allow for sufficient discrimination between two- and three-pitch states; even with the number of voiced sources given, tracking three pitches proved difficult due to the small number of frequency channels dominated by the weakest source.

Two additional issues should be addressed in future work. First, due to the binaural simulation methods used, we consider an azimuth grid with 5° resolution. Of course in practice, source azimuths are not constrained to this grid and further testing is needed to analyze performance for source azimuths that fall between the trained angles. Second, in spite of constraining the search space of the HMM using separate pitch and azimuth modules, computational complexity may be a concern for certain applications. Our current implementation of the HMM in Matlab running on a Dell PowerEdge R710 takes roughly 90 s per 1 s of audio, although no effort has been made to optimize for speed. We are confident that this time could be reduced substantially with an improved implementation, however, a more efficient algorithm for computing the multisource posterior density would likely be necessary in order to achieve real-time processing.

Our long-term goal is the development of a robust binaural system that benefits from but does not rely on spatial cues. While the proposed HMM framework is a step toward that goal, the system still fundamentally relies on azimuth to achieve sequential organization. An interesting direction for future work is thus developing a similar integration of monaural and spatial cues for sequential organization.

ACKNOWLEDGMENT

The authors thank two anonymous reviewers for their constructive criticisms and suggestions. The authors also thank M. Mandel and N. Q. K. Duong for making implementations of their algorithms publicly available, and C. Hummersone for making the set of measured impulse responses available.

REFERENCES

[1] P. Boersma, "Accurate short-time analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," *Inst. Phonetic Sci.*, vol. 17, pp. 97–110, 1993.

[2] J. S. Bradley, "Predictors of speech intelligibility in rooms," *J. Acoust. Soc. Amer.*, vol. 80, pp. 837–845, 1986.

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

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

[5] H. Buchner, R. Aichner, and W. Kellermann, "Blind source separation for convolutive mixtures: A unified treatment," in *Audio Signal Processing for Next-Generation Multimedia Communications Systems*, Y. Huang and J. Benesty, Eds. Dordrecht, The Netherlands: Kluwer, 2004, pp. 255–294.

[6] D. R. Campbell, "The ROOMSIM User Guide (v3.3)," 2004 [Online]. Available: http://media.paisley.ac.uk/~campbell/Roomsim/

[7] 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.

[8] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1830–1840, Sep. 2010.

[9] A. S. Feng and D. L. Jones, "Location-based grouping," in *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, D. L. Wang and G. J. Brown, Eds. New York: Wiley/IEEE Press, 2006, pp. 187–208.

[10] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, no. 8, pp. 926–935, Aug. 1972.

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

[12] 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

[13] D. J. Hand, "Idiot's Bayes—Not so stupid after all?," *Int. Statist. Rev.*, vol. 69, no. 3, pp. 385–398, 2001.

[14] R. C. Hendriks, R. Heusdens, U. Kjems, and J. Jensen, "On optimal multichannel mean-squared error estimators for speech enhancement," *IEEE Signal Process. Lett.*, vol. 16, no. 10, pp. 885–888, Oct. 2009.

[15] 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.

[16] J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Joint DOA and fundamental frequency estimation methods based on 2-d filtering," in *Proc. EUSIPCO*, 2010.

[17] 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.

[18] 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.

[19] M. Képesi, F. Pernkopf, and M. Wohlmayr, "Joint position pitch tracking for 2-channel audio," in *Proc. Int. Workshop Content Based Multimedia Indexing*, 2007.

[20] G. Kim, Y. Lu, Y. Hu, and P. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *J. Acoust. Soc. Amer.*, vol. 126, no. 3, pp. 1486–1494, 2009.

[21] R. F. Lyon, "A computational model of binaural localization and separation," in *Proc. ICASSP*, 1983, pp. 1148–1151.

[22] N. Ma, J. Barker, H. Christensen, and P. Green, "Binaural cues for fragment-based speech recognition in reverberant multisource environments," in *Proc. INTERSPEECH*, 2011.

[23] 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.

[24] T. Nakatani, M. Goto, and H. G. Okuno, "Localization by harmonic structure and its application to harmonic sound stream segregation," in *Proc. ICASSP*, 1996, pp. 653–656.

[25] 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.

[26] 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.

[27] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An Efficient Auditory Filterbank based on the Gammatone Function, MRC App. Psych. Unit," Tech. Rep. Cambridge, 1988.

[28] N. Roman, D. L. Wang, and G. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Amer.*, vol. 114, pp. 2236–2252, 2003.

[29] N. Roman and J. Woodruff, "Intelligibility of reverberant noisy speech with ideal binary masking," *J. Acoust. Soc. Amer.*, vol. 130, pp. 2153–2161, 2011.

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

[31] B. G. Shinn-Cunningham, "Influences of spatial cues on grouping and understanding sound," in *Proc. Forum Acusticum*, 2005.

[32] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays: Signal Processing Techniques and Applications*. New York: Springer, 2001, pp. 39–60.

[33] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 260–275, Feb. 2010.

[34] 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.

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

[36] D. L. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner, "Speech intelligibility in background noise with ideal binary time-frequency masking," *J. Acoust. Soc. Amer.*, vol. 125, pp. 2336–2347, 2009.

[37] R. Weiss, M. Mandel, and Ellis, "Combining localization cues and source model constraints for binaural source separation," *Speech Commun.*, vol. 53, pp. 606–621, 2011.

[38] 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.

[39] J. Woodruff and D. L. Wang, "Binaural localization of multiple sources in reverberant and noisy environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 5, pp. 1503–1512, Jul. 2012.

[40] W. S. Woods, M. Hansen, T. Wittkop, and B. Kollmeier, "A simple architecture for using multiple cues in sound separation," in *Proc. ICSLP*, 1996.

[41] S. N. Wrigley and G. J. Brown, "Binaural speech separation using recurrent timing neural networks for joint F0-localisation estimation," in *Machine Learning for Multimodel Interaction*. Berlin/Heidelberg, Germany: Springer, 2008, pp. 271–282.

[42] M. Wu, D. L. Wang, and G. J. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 3, pp. 229–241, May 2003.

[43] J. X. Zhang, M. G. Christensen, S. H. Jensen, and M. Moonen, "Joint DOA and multi-pitch estimation based on subspace techniques," *EURASIP J. Adv. Signal Process.*, vol. 2012, pp. 1–11, 2012.

John Woodruff (S'09–M'12) 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, the M.Mus. degree in music technology from Northwestern University, Evanston, IL, in 2006, and the Ph.D. degree in computer science and engineering from The Ohio State University, Columbus, OH, in 2012.

His research interests include computational auditory scene analysis, music and speech processing, auditory perception and statistical learning. Since 2012, he has been with Audience, Inc.

DeLiang Wang, photograph and biography not provided at the time of publication.