Technical Report OSU-CISRC-4/06-TR44
Department of Computer Science and Engineering
The Ohio State University
Columbus, OH 43210-1277

Ftpsite: **ftp.cse.ohio-state.edu**
Login: **anonymous**
Directory: **pub/tech-report/2006**
File: **TR44.pdf**
Website: **http://www.cse.ohio-state.edu/research/techReport.shtml**

Binaural Tracking of Multiple Moving Sources

Nicoleta Roman

Department of Mathematics

The Ohio State University at Lima, Lima, OH 45805, USA

roman.45@osu.edu

DeLiang Wang

Department of Computer Science and Engineering & Center for Cognitive Science

The Ohio State University, Columbus, OH 43210, USA

dwang@cse.ohio-state.edu

Correspondence should be directed to D.L. Wang: Department of Computer Science and Engineering, The Ohio State University, 2015 Neil Avenue, Columbus, OH 43210.

Phone:(614)-292-6827, URL: http://www.cse.ohio-state.edu/~dwang/

Abstract—This paper addresses the problem of tracking multiple moving sources using binaural input. We observe that binaural cues are strongly correlated with source locations in time-frequency regions dominated by only one source. Based on this observation, we propose a novel tracking algorithm that integrates probabilities across reliable frequency channels in order to produce a likelihood function in the target space, which describes the azimuths of active sources at a particular time frame. Finally, a hidden Markov model (HMM) is employed to form continuous tracks and automatically detect the number of active sources across time. Experimental results are presented for two- and three-source scenarios. A comparison shows that our HMM model outperforms a Kalman filter based approach in tracking active sources across time. Our study represents a first step in addressing auditory scene analysis with moving sound sources.

Index Terms—binaural processing, hidden Markov model (HMM), moving source tracking, multi-source tracking

I. INTRODUCTION

The problem of tracking multiple moving targets arises in many domains including surveillance, navigation and speech processing. In this study we are interested in localizing and tracking multiple acoustic sources that may move, such as concurrent speakers at a cocktail party. A solution to this problem is needed in many speech processing applications such as meeting segmentation, hands-free speech acquisition and hearing prosthesis [1] [2].

Numerous multitarget tracking algorithms have been developed, mostly for radar sensors (for a review see [3]). There are two main approaches to target tracking that utilize Bayesian inference: Multiple hypothesis tracking (MHT) and Bayesian filtering. The MHT attempts to optimally associate the noisy measurements over time to form multiple tracks. For a particular hypothesis, a Kalman filter is associated with each track and a maximum *a posteriori* (MAP) cost is computed using the Kalman filter innovation sequence and the *a priori* track set probability. Finally, the estimated tracks are obtained by comparing all the hypothesized track sets using the MAP cost. Bayesian filtering, on the other hand, aims at the conditional mean estimation of the location state space. The conditional probability is recursively estimated by combining a model for the source motions and a likelihood for the state space given a set of noisy measurements. The Bayesian tracker has a closed-form solution only for a linear process with Gaussian noise which is equivalent to the Kalman filter in this case. In general, optimum MHT and Bayesian solutions require an exponential number of evaluations and therefore are deemed impractical [4]. Hypothesis pruning and merging techniques have been proposed to reduce this computational burden, including measurement gating [5], probabilistic data association [6], and Viterbi based algorithms [7]. An approximation to Bayesian filtering for nonlinear functions, non-Gaussian noises, and multi-modal distributions is provided using sequential Monte-Carlo methods, also known as particle filtering [8] [9]. When the number of active sources rapidly varies the above algorithms require complex birth/death rules to initiate and terminate individual tracks.

HMM has also been proposed for target tracking in sonar networks by employing the Markovian modeling of source dynamics in a discretized target space [10]. It is important to note that this framework can handle multi-modal likelihood distributions. Due to discrete Markov modeling, Viterbi decoding can be used to efficiently search for the most likely state sequences. The number of targets is, however, decided in this algorithm in a postprocessing step based on detection of local maxima in the likelihood distribution.

Several of the above techniques have been adapted and applied to the problem of speaker tracking using microphone arrays. To estimate the locations of active sources in each time frame, these algorithms typically employ variants of the well-known generalized cross-correlation function [11] or subspace-based methods [12]. The particle filtering theory, for example, has been extended to the tracking of one moving speaker in a reverberant environment [13] [14]. For the tracking of multiple speakers, algorithms have been proposed that combine Kalman filtering with probabilistic data association techniques [15] [16]. These multi-source tracking algorithms have been shown to provide good localization results using an array of microphones. However, when restricting the size of the array to only two sensors, as in the case of human audition, the multi-source tracking problem becomes more challenging and little has been attained in this respect. As a solution, visual and auditory information are jointly used for the task, where audition helps mainly in resolving ambiguities during occlusions [17].

Location has been shown to be an effective cue for computational systems that attempt to separate individual talkers in noisy environments using only two microphones [18] [19]. The binaural cues of interaural time differences (ITD) and interaural intensity differences (IID) are strongly correlated with the source locations in time-frequency (T-F) regions dominated by only one source. Hence, with accurate locations, the binaural cues can be used to segregate the original signals. However, in a realistic environment source motion and head movement have to be considered and location estimates may have to be updated every frame of data.

In this paper, we study the tracking of multiple speakers based on the binaural response of a KEMAR dummy head that accurately simulates the filtering process of the head, torso and external ear [20]. We propose a novel HMM framework where the change in the number of active tracks is modeled probabilistically. Specifically, the target space is modeled as a set of subspaces with jump probabilities between them. Each subspace models the tracking of a subset of possible active sources. Hence, unlike previous methods, the detection of tracks in the HMM is fully automatic and does not require heuristic rules for track initialization and termination. Our approach extends an HMM-based model for multi-pitch tracking proposed by Wu et al. [21] [22]. Due to the sparsity of speech signal distribution in a two-dimensional (2-D) T-F representation [23], while some T-F units in a mixture signal respond to overlapping multiple sources, others are dominated by only one source and thus provide reliable information for localization. In this paper, the T-F decomposition is obtained at the output of an auditory filterbank; the output of each filter channel is divided in 20-ms sections with 10-ms overlap that correspond to T-F units. Because the binaural cues are strongly correlated with source locations in the regions dominated by a single source, peaky statistical distributions characterize the observations in the reliable frequency channels. Hence, we propose to use a channel selection mechanism to determine the reliable channels followed by a statistical integration of these channels in order to obtain the likelihood function for different target subspaces.

The rest of the paper is organized as follows: the next section gives an overview of the system. Section III describes auditory motion modeling. Section IV briefly describes the auditory periphery model and binaural processing. Section V contains details of the proposed statistical model. In this paper we report experimental results for the tracking of two and three simultaneous speakers. Section VI gives the simulation results and a comparison with a Kalman filter approach. The last section concludes the paper.

II. MODEL ARCHITECTURE

Our multi-source tracking system consists of the following four stages: 1) a model of the auditory periphery and binaural cue estimation; 2) a channel selection mechanism that identifies reliable frequency channels in each time frame; 3) a multichannel statistical integration method that produces the likelihood function for target subspaces; and 4) a continuous HMM model for multi-source tracking. Fig. 1 illustrates the model architecture for the case of two moving sources.

The input to our model is a binaural response of a KEMAR dummy head to an acoustic scene with multiple moving sources. We utilize here the catalog of head related transfer functions (HRTF) measured by Gardner and Martin [24] for anechoic conditions at fixed source locations on a sphere around the KEMAR. Interpolation is then used to obtain HRTF responses for arbitrary positions on the sphere. HRTFs introduce a natural combination of ITD and IID into the

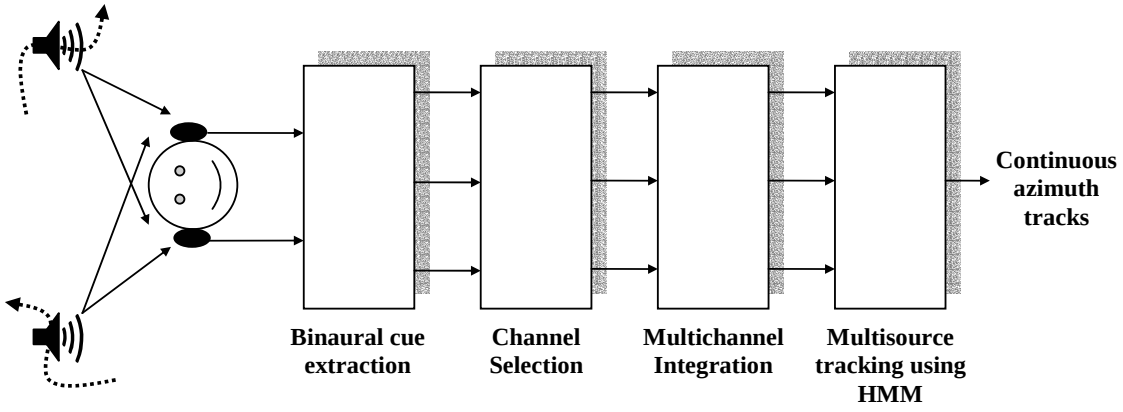

Fig. 1. A schematic diagram of the proposed multi-source tracking system.

signals which is extracted in subsequent stages of our model. Here we restrict the motion of individual sources to the half horizontal plane with azimuth in the range [-90°, 90°]. The system is, however, extensible to cover the entire azimuth range since ITD and IID used jointly can potentially differentiate between the front and the back. Hence, for each moving source left and right ear signals are obtained by filtering with time-varying HRTFs that correspond to a source trajectory on the frontal semicircle. The responses to multiple sources are added at the two ears and form the binaural input to our system.

In the first stage, the resulting left and right ear mixtures are analyzed using an auditory periphery model. Then, for each frequency channel, normalized cross-correlation functions between the two ear signals are computed in consecutive time frames. The time lag of a peak in the cross-correlation function is a candidate for ITD estimation. At high frequencies multiple peaks are present and this creates ambiguity in localization. We resolve this ambiguity by using IID information.

Channel selection comprises the second stage of our system. This stage attempts to select reliable channels defined as those dominated primarily by only one source while removing the more corrupted ones. Here, we use the height of the peak in the cross-correlation function as a measure of channel reliability. The third stage is the multichannel integration of location information. The conventional approach is to summate the cross-correlation functions across all frequency channels [18]. A peak in the summary cross-correlation suggests an active source while the height of the peak indicates its likelihood. This approach, however, under-utilizes the location information in individual frequency channels. In our system, we consider the statistical distribution of the ITD-IID estimates. Given a configuration hypothesis, we first formulate the probability of each channel supporting the hypothesis and then employ an integration method to produce the likelihood of observing the configuration. For configurations with more than one active source a gating mechanism is used to associate the observations with one of the sources.

The last stage of the algorithm is to form azimuth tracks in a continuous HMM framework. We propose an HMM model that allows jumping between subspaces within each of which only a subset of the total number of sources is active. The framework combines the likelihood model from the previous stage, a model for the dynamics of source motion and jump probabilities

between the individual subspaces. Finally, optimal azimuth tracks are obtained using the Viterbi decoding algorithm.

III. MODELING AUDITORY MOTION

For human audition, sound source localization is primarily achieved with the binaural cues of ITD and IID. For a moving sound, there are changes in ITD and IID that may provide velocity information and enable the listener to perceive and track the changing source location [25]. The transmission path between the acoustic source and the receiver contains many subsystems, i.e. the loudspeaker, the ear canal and the eardrum (microphone). Here, we use the diffuse-field equalized HRTFs for which all the factors that are not location-dependent are eliminated. The HRTF catalog [24] provides 256 point impulse responses for a fixed number of locations residing on a 1.4 m radius sphere around the KEMAR head. In particular, the resolution in the horizontal plane is 5^o azimuth. The sampling rate is fixed at 44.1 kHz.

An attractive property of HRTFs is that they are almost minimum-phase [26]. Therefore, a standard way of modeling HRTFs is to decompose the system into a cascade of a minimum-phase filter and a pure delay line [27]. The motivation is that minimum-phase systems behave better than the raw measurements for interpolation both in the phase and the magnitude response. In addition, a minimum-phase reconstruction of HRTF does not have perceptual alterations [28]. Here, we reconstruct the minimum-phase part through appropriate windowing in the cepstral domain. Specifically, the negative cepstral coefficients are set to 0 and a minimum-phase filter is then obtained by inverting the truncated cepstrum [29]. The time delay part is estimated as the mean of the group delay in the range of interest from 80 Hz to 5 kHz.

To simulate a continuous motion, the impulse response of an arbitrary direction of sound incidence is obtained by interpolating separately the minimum-phase filters and the time delays corresponding to neighboring entries in the HRTF catalog. Since we simulate motions in the horizontal plane, a simple two-way linear interpolation is applied. The impulse response is then reconstructed from the cascade of the resulting minimum-phase filter and the time delay. Finally, to synthesize the binaural response of the KEMAR dummy head to one moving source a monaural signal is upsampled to 44.1 kHz and filtered with the corresponding time-varying left and right impulse responses. The synthesized multiple sources are added at the two ears and fed to the tracking system.

IV. AUDITORY PERIPHERY AND BINAURAL PROCESSING

It is widely acknowledged that cochlear filtering can be modeled by a bandpass filterbank [30]. The filterbank employed here consists of 128 fourth-order gammatone filters [31] with channel center frequencies equally distributed on the equivalent rectangular bandwidth (ERB) scale between 80 Hz and 5 kHz. In addition, we adjust the gains of the gammatone filters in order to simulate the middle ear transfer function [32]. In the final step of the peripheral model, we use a simple model of hair cell transduction that consists of half-wave rectification and a square root operation.

To extract ITD information, we employ the normalized cross-correlation computed at lags equally distributed from –1 ms to 1 ms $(-44 < \tau < 44)$ using a rectangular integration window of

20 ms (corresponding to *K*=880 samples below). This range of time lags encloses the plausible range for the human head. The cross-correlation is computed for all frequency channels and updated every 10 ms, according to the following formula for frequency channel *c*, time frame *m*, and lag τ:

$$C(c,m,\tau)=\frac{\sum_{k=0}^{K-1}(l_c(m-k)-\overline{l}_c)(r_c(m-k-\tau)-\overline{r}_c)}{\sqrt{\sum_{k=0}^{K-1}(l_c(m-k)-\overline{l}_c)^2}\sqrt{\sum_{k=0}^{K-1}(r_c(m-k-\tau)-\overline{r}_c)^2}}, \quad (1)$$

where l_c, r_c refer to the left and right peripheral output for channel *c*, and $\overline{l}_c$, $\overline{r}_c$ their mean values over the integration window, respectively. Each lag τ corresponding to a peak in the cross-correlation function is considered an ITD estimate. In addition, IID information is extracted for frequency channel *c* and time frame *m* by computing the energy ratio at the two ears, expressed in decibels:

$$\iota=20\log_{10}\left(\sum_{k=0}^{K-1}r_c^2(m-k)\Big/\sum_{k=0}^{K-1}l_c^2(m-k)\right). \quad (2)$$

V. STATISTICAL TRACKING

The problem of tracking the azimuths of multiple acoustic sources is formulated here in an HMM framework. An HMM is a doubly stochastic process where an underlying stochastic (Markovian) process that is not directly observable (i.e. "hidden") is observed through another stochastic process that produces a sequence of observations [33]. An HMM is completely defined by the following: 1) the possible target state space; 2) the transition probabilities that reflect the evolution of the target states across time; and 3) the observation probabilities conditioned on the target states, also known as the observation likelihood. Fig. 2 illustrates our proposed HMM framework. A state in the target space specifies what the active sources are as well as their azimuth information at a particular time frame. The target space is decomposed into subspaces; each subspace corresponds to a subset of active sources. Hence, the transition probability between states in neighboring time frames must take into account both the jump probability between subspaces and the temporal evolution within individual subspaces. Finally, a statistical model that integrates ITD and IID observations in different frequency channels is used to construct the observation likelihood in the target space. To increase the robustness of the system only frequency channels that are dominated by a single source and thus deemed reliable are considered in our statistical integration.

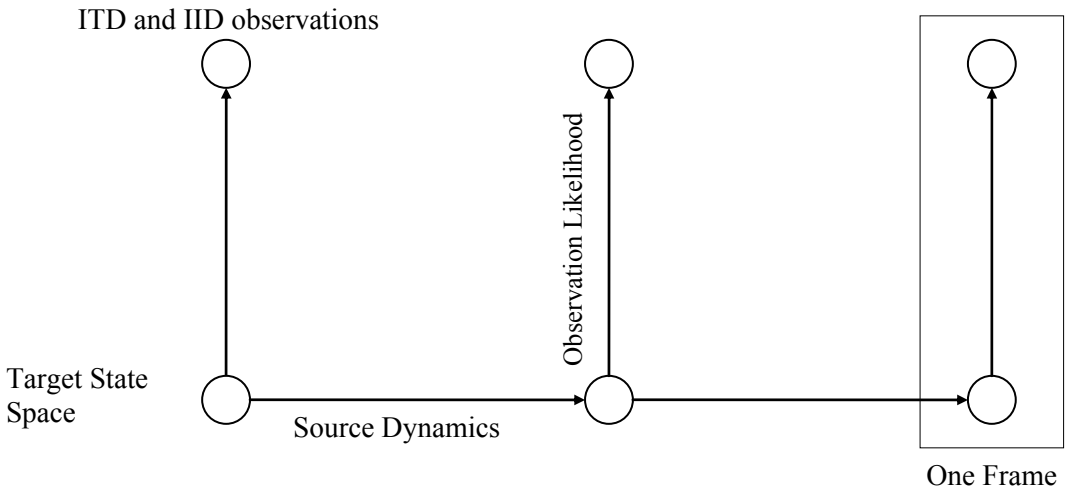

Fig. 2. Schematic diagram of an HMM for modeling continuous source tracks.

A. Dynamics Model

In a practical multi-source tracking situation, the number of active sources at a particular time is generally unknown. In this study, we assume a maximum of three sources and aim to assign separate tracks to each of the sources; the framework can be extended for more sources. Hence, we define the target state space as the union of eight possible subspaces as follows:

$$S = S_0 \bigcup S_1^1 \bigcup S_1^2 \bigcup S_1^3 \bigcup S_2^{1,2} \bigcup S_2^{1,3} \bigcup S_2^{2,3} \bigcup S_3, \tag{3}$$

where S_0 is the silence space with no active source, S_1^i is the state space for a single active source *i*, $S_2^{i,j}$ is the state space for two simultaneously active sources *i* and *j*, and S_3 is the state space for all three active sources. A state is represented as a 3-D vector $\boldsymbol{x} = (\varphi^1, \varphi^2, \varphi^3)$, where each dimension φ^i gives the azimuth for the *i*th source or indicates that the source is silent.

State transitions in a Markov model provide a standard statistical framework for dealing with multiple dynamic models (e.g. [4]). Suppose that the state of the system at frame *m*, $\boldsymbol{x}_m = (\varphi_m^1, \varphi_m^2, \varphi_m^3)$, is in the subspace s_m and the sources are independent of each other. Then the state transitions are described by:

$$p\left(\boldsymbol{x}_m, s_m \mid \boldsymbol{x}_{m-1}, s_{m-1}\right) = p(s_m \mid s_{m-1}) \prod_{i \in I} p\left(\varphi_m^i \mid \varphi_{m-1}^i\right), \tag{4}$$

where $p(s_m \mid s_{m\text{-}1})$ is the jump probability between subspaces, *I* is the set of active sources at time frame *m*, and $p\left(\varphi_m^i \mid \varphi_{m-1}^i\right)$ gives the temporal evolution of the *i*th source.

TABLE I

JUMP PROBABILITIES BETWEEN SUBSPACES WITH ZERO, ONE, TWO AND THREE ACTIVE SOURCES

	$\rightarrow S_0$	$\rightarrow S_1^1$	$\rightarrow S_1^2$	$\rightarrow S_1^3$	$\rightarrow S_2^{1,2}$	$\rightarrow S_2^{1,3}$	$\rightarrow S_2^{2,3}$	$\rightarrow S_3$
S_0	0.9663	0.0112	0.0112	0.0112	0	0	0	0
S_1^1	0.0692	0.6590	0	0	0.1359	0.1359	0	0
S_1^2	0.0692	0	0.6590	0	0.1359	0	0.1359	0
S_1^3	0.0692	0	0	0.6590	0	0.1359	0.1359	0
$S_2^{1,2}$	0	0.0347	0.0347	0	0.7077	0	0	0.2230
$S_2^{1,3}$	0	0.0347	0	0.0347	0	0.7077	0	0.2230
$S_2^{2,3}$	0	0	0.0347	0.0347	0	0	0.7077	0.2230
S_3	0	0	0	0	0.0448	0.0448	0.0448	0.8655

The jump probabilities between state spaces of zero-, one-, two- and three-sources in consecutive time frames are estimated using mixtures of three speech utterances from the TIMIT database [34]. For this, speech activity detection is performed separately on each individual utterance by using a threshold on the signal energy. This enables the detection of the number of active sources at each time frame in the mixture. We assume that at most one source can be turned on or off during one time frame. Also, the three one-source as well as the three two-source subspaces are considered equally probable. The resulting jump probabilities between the eight subspaces are reported in Table I.

We assume that an active source moves slowly and follows a linear trajectory with additive Gaussian noise. Also, when a source transitions from silence to activity we assume a uniform distribution in the azimuth space. Therefore the dynamics of the *i*th source is described by:

$$p\left(\varphi_m^i \mid \varphi_{m-1}^i\right) = \begin{cases} N(\varphi_{m-1}^i, \sigma), & \varphi_{m-1}^i \neq nil \\ U(\varphi_m^i), & \varphi_{m-1}^i = nil \end{cases}, \tag{5}$$

where *nil* stands for silence, $N(\varphi, \sigma)$ denotes the Gaussian distribution with mean φ and standard deviation σ which is set to a small value. U denotes the uniform distribution in the azimuth range [-90°, 90°].

B. Statistics of ITD and IID

For a particular T-F unit, the normalized cross-correlation function of (1) has a maximum of 1 when the left and right signals are identical except for a time delay and an intensity difference. This condition is satisfied when only one source is active in the corresponding T-F unit. The computed ITD and IID reflect in this case the actual source location. However, when sources from different locations are all strong in a T-F unit, the left and right mixtures do not satisfy this condition anymore and the maximum in the normalized cross-correlation function decreases.

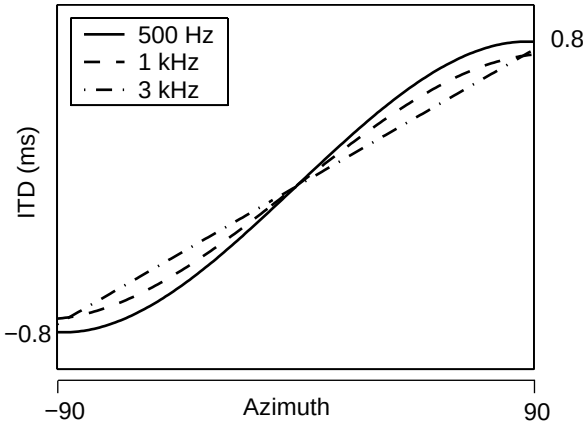

Fig. 3. ITD reference functions for three auditory channels with center frequencies of 500 Hz, 1 kHz, and 3 kHz and azimuth in the range [-90°, 90°].

Moreover, ITD and IID deviate from the actual source locations and can indicate phantom sources [18]. Hence, we utilize the peak height of the cross-correlation function as a measure of reliability in individual T-F units: A T-F unit is considered reliable (i.e., dominated by only one source) and thus selected if its peak height exceeds a threshold $\theta(c)$. The thresholds $\theta(c)$ are estimated so that 80% of all noisy T-F units are rejected. A unit is considered noisy if the relative strength *R* between target signal and interference is less than 0.2 where *R* is defined as the ratio between target energy and the sum of target and interference energy. We observe that $\theta(c)$ is a linearly decreasing function with respect to channel index *c*.

For each selected T-F unit, the estimated ITD and IID signal a specific source location. By studying the deviation of the estimated ITD and IID values from the reference values, we can derive the probability of one selected channel supporting a location hypothesis. For each frequency channel, the reference values are obtained from simulated white noise signals at locations in the azimuth range [-90°, 90°]. Fig. 3 shows ITD values for three auditory channels with center frequncies of 500 Hz, 1 kHz and 3 kHz where the ITD corresponds to the lag of the maximum peak in the cross-correlation function. As seen in the figure, ITD is monotonic with respect to azimuth but has a slight dependency on channel center frequency due to diffraction effects [35]. IID reference values for all frequency channels are also shown in Fig. 4. Note that IID is highly dependent on both channel frequency and azimuth.

Consider channel *c* and azimuth φ for which the ITD and IID reference values are $\tau_{ref}(c,\varphi)$ and $\iota_{ref}(c,\varphi)$. For a given T-F unit, we define the ITD and IID deviations as:

$$\delta_\tau = \tau - \tau_{ref}(c,\varphi), \tag{6a}$$

$$\delta_\iota = \iota - \iota_{ref}(c,\varphi), \tag{6b}$$

where τ is the lag of the closest peak in the cross-correlation function to $\tau_{ref}(c,\varphi)$ and ι is the estimated IID.

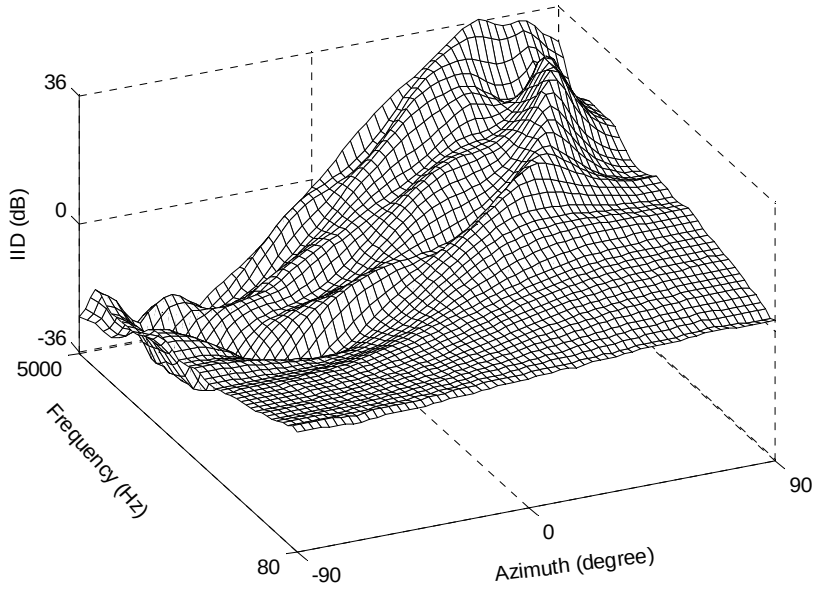

Fig. 4. IID reference functions for frequency in the range 80 Hz – 5000 Hz and azimuth in the range [-90°, 90°].

Statistics of the deviations δ_τ and δ_ι are collected separately for each frequency channel across different time frames. Fig. 5 shows the results of these deviations for a channel with center frequency f_c of 1.5 kHz. The ITD and IID deviations are obtained for the one-source scenario using a small set of 10 utterances from the TIMIT database and various linear motion patterns. As seen in the figure, both histograms are centered at zero and decrease sharply on both sides of zero. Consequently, we model the joint distribution of ITD and IID deviations in channel *c* as a combination of a Laplacian distribution, and a uniform distribution which models the background noise:

$$p_c(\delta_\tau, \delta_\iota) = (1-q)L(\delta_\tau, \lambda_\tau(c))L(\delta_\iota, \lambda_\iota(c)) + qU_c(\Delta_\tau, \Delta_\iota), \tag{7}$$

where $0 < q < 1$ is the noise level. $U_c(\Delta_\tau, \Delta_\iota)$ is the 2-D uniform distribution in the plausible range for $\delta_\tau \in [-\Delta_\tau, \Delta_\tau]$ in lag step and $\delta_\iota \in [-\Delta_\iota, \Delta_\iota]$ in dB. $\Delta_\iota = 20$ and $\Delta_\tau = \max(\frac{f_s}{2f_c}, 44)$, where f_s is the sampling frequency and 44 lag steps correspond to a delay of 1 ms. $L(\delta, \lambda)$ is the Laplacian distribution with parameter λ defined by:

$$L(\delta, \lambda) = \frac{1}{2\lambda} \exp\left(-\frac{|\delta|}{\lambda}\right). \tag{8}$$

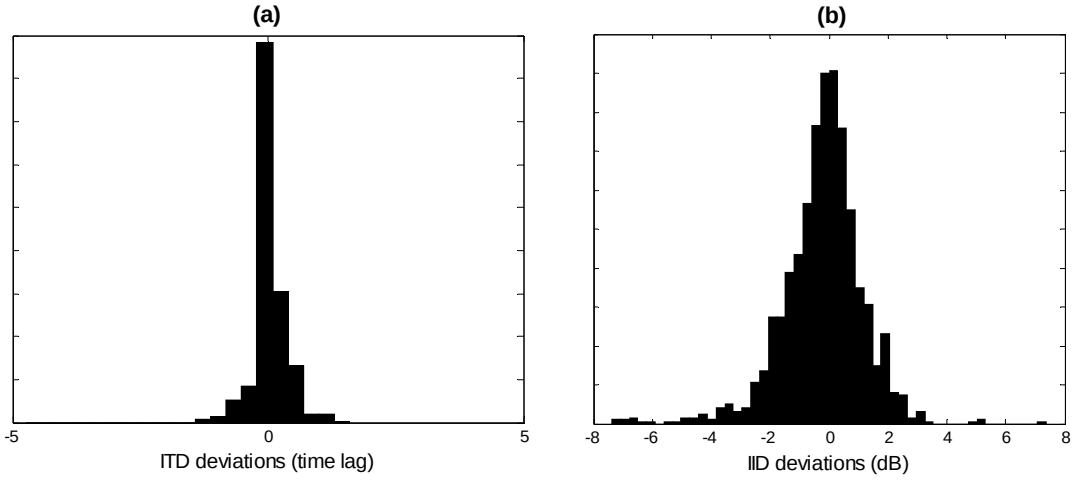

Fig. 5. Histogram of estimated ITD and IID deviations from reference values for a channel with f_c = 1.5 kHz in the one-source scenario.

We observe that the parameters $\lambda_\tau(c)$, $\lambda_\iota(c)$ are channel dependent: $\lambda_\tau(c)$ decreases abruptly with increasing c (or f_c) whereas $\lambda_\iota(c)$ increases slowly. To obtain smooth parameters across channels we use the following simple approximation:

$$\lambda_\tau(c) = a_1 + a_2 / f_c, \tag{9a}$$

$$\lambda_\iota(c) = a_3 + a_4 \cdot c. \tag{9b}$$

Similarly, ITD and IID statistics are extracted for multi-source scenarios with two and three active sources. We employ a set of 10 binaural mixtures using the same utterances as in the one-source situation and various linear motion patterns. For a selected T-F unit, the dominant source is obtained by comparing the energies of the individual sources and the ITD and IID deviations are computed relative to the dominant source. While the deviations exhibit the same peaky distributions as in the one-source scenario, their variance increases due to the mutual interference between the sources.

The maximum likelihood (ML) method is then used to estimate the parameters a_1, a_2, a_3, and a_4 for the one-source and the multi-source scenarios assuming a fixed noise level q across all conditions and frequency channels. This ensures that the background noise and the unreliable channels do not influence the comparison between one-source and multi-source scenarios. ML estimation gives q=0.03. The parameters a_1, a_2, a_3, and a_4 are reported in Table II.

TABLE II
ESTIMATED MODEL PARAMETERS FOR ONE-SOURCE AND MULTI-SOURCE CONDITIONS

	a_1	a_2	a_3	a_4
One-source	0.1328	59.0497	0.3666	0.0026
Multi-source	0.1293	500.000	1.2306	0.0071

C. Likelihood Model

In this subsection we derive the conditional probability density $p(\{T_c, \iota_c\} \mid \boldsymbol{x})$, often referred to as the likelihood, which statistically describes what a single frame of ITD and IID observations relate to the joint state $\boldsymbol{x}$ of the source locations to be tracked. Here, T_c is the set of time lags τ_c corresponding to the local peaks in the cross-correlation function and ι_c is the estimated IID for channel *c*. The braces denote all frequency channels.

First, we consider the conditional probability $p(\{T_c, \iota_c\} \mid \boldsymbol{x})$ for the one-source subspaces, i.e. $\boldsymbol{x} \in S_1^1 \bigcup S_1^2 \bigcup S_1^3$. For channel *c*, we compute the deviations δ_τ, δ_ι as described in Eq. 6 using as reference values $\tau_{ref}(c,\varphi)$ and $\iota_{ref}(c,\varphi)$ where φ refers to the azimuth of the hypothesized active source. Then, the conditional probability of the observations in channel *c* with respect to the one-source state $\boldsymbol{x}$ is given by:

$$p(T_c, \iota_c \mid \boldsymbol{x}) = \begin{cases} p_c(\delta_\tau, \delta_\iota), & \textit{if channel c is selected} \\ qU_c(\Delta_\tau, \Delta_\iota), & \textit{else} \end{cases}, \tag{10}$$

where the symbols are as described in Eq. 7 and Eq. 9 and the parameters are estimated for the one-source scenario. Note that the uniform background noise is assigned to an unreliable channel.

By assuming independence between observations in different channels, the conditional probability in a frame can be easily obtained by multiplying the conditional probabilities in individual channels. However, the observations are usually correlated due to the wideband nature of speech signals and the overlapping passbands of neighboring gammatone filters. This correlation results in 'spiky' distributions. This is known as the probability overshoot phenomenon. To alleviate this problem, the observation probability in the current time frame conditioned on the one-source state $\boldsymbol{x}$ is smoothed using a root operation [36]:

$$p(\{T_c, \iota_c\} \mid \boldsymbol{x}) = \kappa \sqrt[N_b]{\prod_c p(T_c, \iota_c \mid \boldsymbol{x})}, \tag{11}$$

where N_b=20 is the root number and κ is a normalization factor.

Next, we consider the conditional probability $p(\{T_c, \iota_c\} \mid \boldsymbol{x})$ for the two-source case, i.e. $\boldsymbol{x} \in S_2^{1,2} \bigcup S_2^{1,3} \bigcup S_2^{2,3}$. Similar to the one-source case, we compute the deviations δ_τ^k and δ_ι^k with respect to the *k*th hypothesized source, where $k = 1,\ 2$. The conditional probability is identical for the three subspaces ($S_2^{1,2}$, $S_2^{1,3}$ and $S_2^{2,3}$) and the *k*th source denotes one of the two active sources in a given subspace. Observe that a selected channel should signal only one source under the assumption that only one speaker dominates a reliable T-F unit. Moreover, all channels whose ITD and IID deviations with respect to the same source are relatively small should support the same source hypothesis. Consequently, we employ a gating technique to associate channels with the hypothesized sources. Specifically, we label channel *c* as belonging to the *k*th source if the corresponding deviations satisfy $\left|\delta_\tau^k\right| < \varepsilon\, \lambda_\tau(c)$ and $\left|\delta_\iota^k\right| < \varepsilon\, \lambda_\iota(c)$ where $\varepsilon = 5$ is the gate size. Assume that the *k*th source is the stronger among the two (most selected channels are dominated by the *k*th source). Then the conditional probability for channel *c* under this assumption is given by:

$$p(T_c, \iota_c \mid \boldsymbol{x}, k) = \begin{cases} qU_c(\Delta_\tau, \Delta_\iota), & \textit{if channel c not selected} \\ p_c(\delta_\tau^k, \delta_\iota^k), & \textit{if channel c belongs to source k}, \\ \max[p_c(\delta_\tau^1, \delta_\iota^1), p_c(\delta_\tau^2, \delta_\iota^2)], & \textit{else} \end{cases} \tag{12}$$

where all the parameters are derived for the multi-source case.

We apply integration of the individual probabilities across all channels as done in Eq. 11 to give the conditional probability $p(\{T_c, \iota_c\} \mid \boldsymbol{x}, k)$ for the current time frame under the assumption that the *k*th hypothesized source is the strongest. Finally, the conditional probability $p(\{T_c, \iota_c\} \mid \boldsymbol{x})$ for the current time frame is the larger of assuming either the first or the second hypothesized source to be the stronger source:

$$p(\{T_c, \iota_c\} \mid \boldsymbol{x}) = \alpha_2 \max[p(\{T_c, \iota_c\} \mid \boldsymbol{x}, 1),\ p(\{T_c, \iota_c\} \mid \boldsymbol{x}, 2)], \tag{13}$$

where α_2 is used to adjust the relative strength of the two-source subspace.

Note that, without the gating mechanism, Eqs. 12 and 13 simplify to a simple max operation in the selected channels. However, this operation tends to overfit the data with a two-source model by assigning the noisy observations produced by one source to two closely spaced sources. The gating mechanism is one way to penalize the overfitting due to noise.

Similar to the two-source case, we consider the conditional probability $p(\{T_c, \iota_c\} \mid \boldsymbol{x})$ for the three-source case, i.e. $\boldsymbol{x} \in S_3$. Eqs. 12 and 13 are easily extensible to three sources by considering all the three-source permutations and utilizing an additional parameter α_3 to adjust the relative strength of the S_3 subspace.

After training we fix α_n as follows: $\alpha_2 = 1$ and $\alpha_3 = e^{-4.25}$. Finally, we fix the probability of the current time frame conditioned on the silence state, i.e. $\boldsymbol{x} \in S_0$:

$$p(\{T_c, t_c\} \mid \mathbf{x}) = \kappa\alpha_0, \tag{14}$$

where $\alpha_0 = e^{-60}$. The above α parameters provide different weights for the individual subspaces. In addition to the actual active sources, a few unreliable channels may align and thus indicate the presence of a spurious source. The differential weights exceed the probability produced by these channels and as a result the system avoids this spurious source occurrence.

D. HMM-Based Source Tracking

For the continuous HMM framework described above, the state space and the time axis are discretized and the standard Viterbi algorithm is employed in order to identify the optimal sequence of states [37]. The algorithm attempts to reconstruct the initial tracks of the most probable sound sources in the scene. Consequently, the decision of the system at every time frame includes the number of currently active sources and their estimated locations.

The computational cost of our HMM framework is mainly due to the large target space which increases with the maximum number of sources considered. This cost can be reduced significantly by employing several efficient implementation techniques. First, the computations are performed in the log domain thus reducing the number of multiplication and root operations. Second, pruning is used to reduce the number of states to be searched for deciding the current candidate states. Since the original tracks move slowly, the difference of azimuths in consecutive time frames, hence search, can be restricted considerably. Specifically, we allow an azimuth range of [-3σ, 3σ] where $\sigma = 2^\circ$ is the standard deviation in the motion model of individual sources. Finally, beam search is employed to reduce the state space considered in the evaluation of the current time frame [38]. In each time frame, beam searching is performed so that any state whose maximum log probability falls more than 20 below the maximum of all states is not considered.

VI. RESULTS AND COMPARISON

The HMM tracking system presented in Section V has been evaluated for two-source and three-source scenarios. As described in Section III, binaural synthesis is used to generate moving sources in the auditory space of a KEMAR dummy head. Given a binaural mixture as input, the system aims at identifying the number of active speakers at a particular time and constructing continuous trajectories for each of the sources.

Fig. 6 shows the result of tracking two simultaneous speakers: one male and one female for a duration of 2.5 s. In this and subsequent evaluations, the original speech utterances are equalized to have the same energy level before binaural synthesis. As seen in the figure, the speakers follow a linear motion with respect to the azimuth on the frontal semicircle. The first speaker moves from 40°, which is on the right side of the KEMAR, to -40° on the left side while the second speaker starts at -40° and ends at 40°. Hence, the two trajectories intersect each other in the middle. The system is able to indicate when a source is active and track the two sources across time as long as it is not entirely masked by the interference. Two types of gaps are detected by the system: when the source is silent and when the source is masked across all

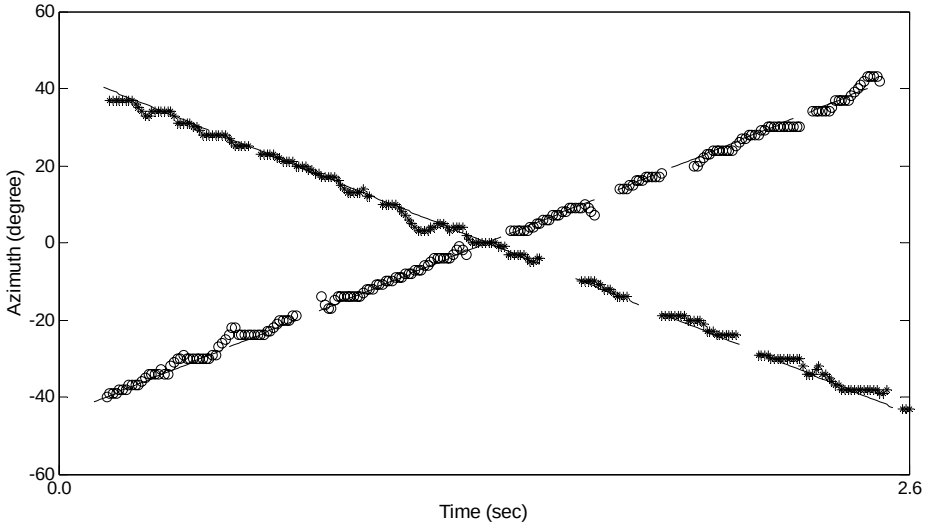

Fig. 6. Source tracking for two crossing sources with linear motion. The solid lines show the true trajectories where a gap indicates a pause in the sentence. The '∗' and 'o' tracks correspond to the estimated tracks.

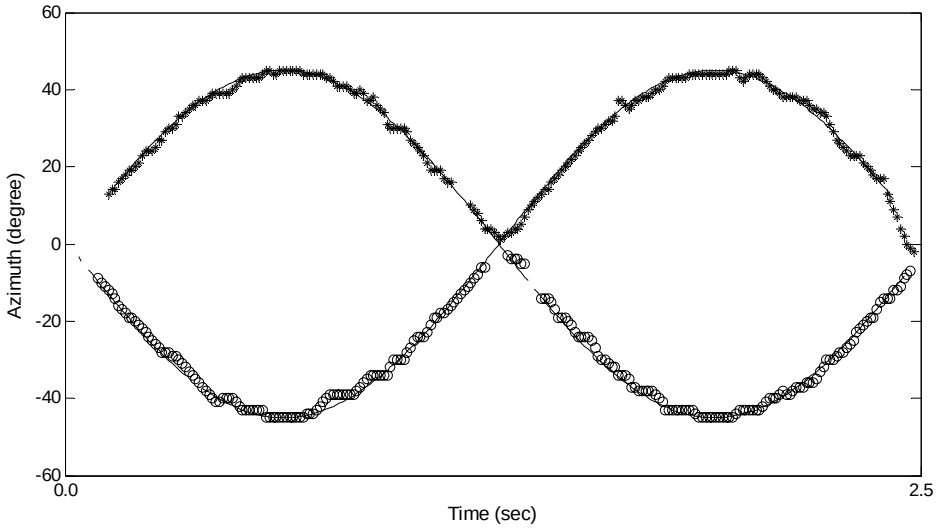

Fig. 7. Source tracking for two crossing sources with nonlinear motion. The solid lines show the true trajectories where a gap indicates a pause in the sentence. The '∗' and 'o' tracks correspond to the estimated tracks.

frequency channels by the other source. While in Fig. 6 the system is able to sequentially link the two sources across the intersection point, in general our system provides no explicit mechanism for disambiguating intersecting source tracks.

Although linear motions have been used during training, our system works for nonlinear motions. Fig. 7 shows the result of tracking one female and one male speaker moving on two cosine azimuth trajectories that also cross each other in the middle. Note that while the two source locations are correctly identified across time, the system switches the trajectories after the

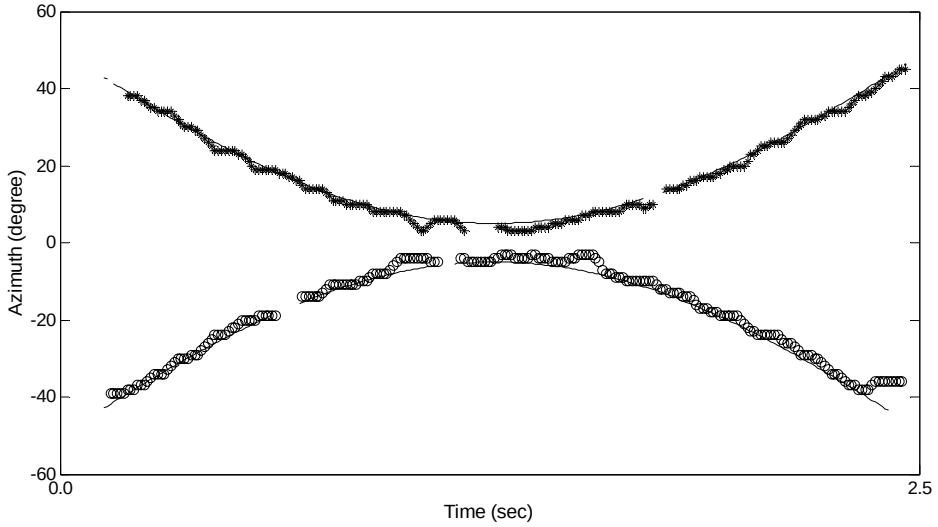

Fig. 8. Source tracking for two sources with closely spaced motions. The solid lines show the true trajectories where a gap indicates a pause in the sentence. The '∗' and 'o' tracks correspond to the estimated tracks.

intersection point. However, as seen in Fig. 6 our system could disambiguate between two tracks at a crossing point when the likelihood is dominated by a single continuous source in the neighborhood of the point. In Fig. 6, the source corresponding to the '○' track is dominated by the source corresponding to the '∗' track around the crossing point, which facilitates the tracking of the latter one and helps the disambiguation of the two tracks.

Fig. 8 highlights the robustness of the system to close trajectories. Two male speakers are moving on nonlinear trajectories with respect to azimuth. The two trajectories are symmetric with respect to the median plane. The first speaker oscillates on the right side of the KEMAR while the second trajectory oscillates on the left side. Note that the distance between the two trajectories can be as small as 10° when both speakers approach the median plane. As seen in the figure, the system makes associations and reconstructs the two trajectories. In some cases, a strong source may mask the presence of other sources, which results in the gaps in the estimated tracks.

Fig. 9 shows results for a challenging scenario with three speakers following nonlinear motions. Two male and one female utterances are used to obtain the three binaural signals. The left ear signal for each speaker is displayed in Fig. 9(a), Fig. 9(b) and Fig. 9(c), respectively. As seen in the figure, the system is able to detect the pauses between words in the utterances. Such word level accuracy is required in real speech applications where the talkers may utter only a few words for the duration of a particular recording. Since we assume that at most one source can be turned on or off during one time frame, there are no transitions allowed between the 1-source subspace and the three-source subspace. In Fig.9, the number of active sources in the time interval [0.45 s, 0.5 s] changes between three sources to one source and then to three sources again. This causes the switching of the tracks corresponding to the first and the third speakers as seen in Fig. 9(d).

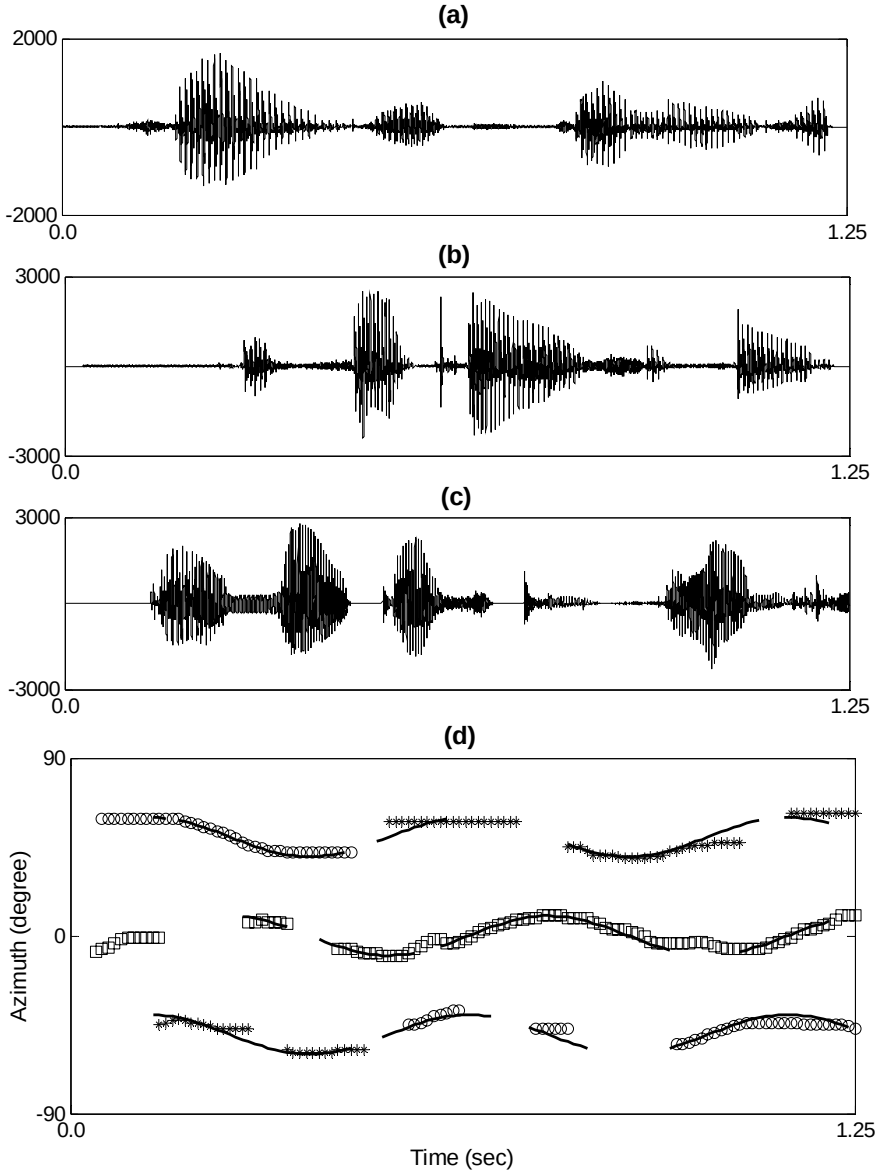

Fig. 9. Tracking three nonstationary moving sources. (a) Left ear signal for the first speaker. (b) Left ear signal for the second speaker. (c) Left ear signal for the third speaker. (d) Continuous tracks obtained by the proposed model. The solid lines show the true trajectories where a gap indicates a pause in the sentence. The '∗', 'o' and '□' tracks correspond to the estimated tracks.

Finally, we compare our approach with a combination of Kalman filtering and data association techniques proposed by Sturim et al. [15] for the tracking of multiple speakers using measurements from an array of 16 microphones. Fig. 10 shows the extracted tracks using this Kalman filtering approach for the same three source configuration as used in Fig. 9. For azimuth estimation, we employ the skeleton cross-correlogram described in [18] which is similar to the generalized cross-correlation method. First, the time-delay axis for the normalized cross-correlations is mapped to the azimuth axis using the reference ITD values. Next, each peak in

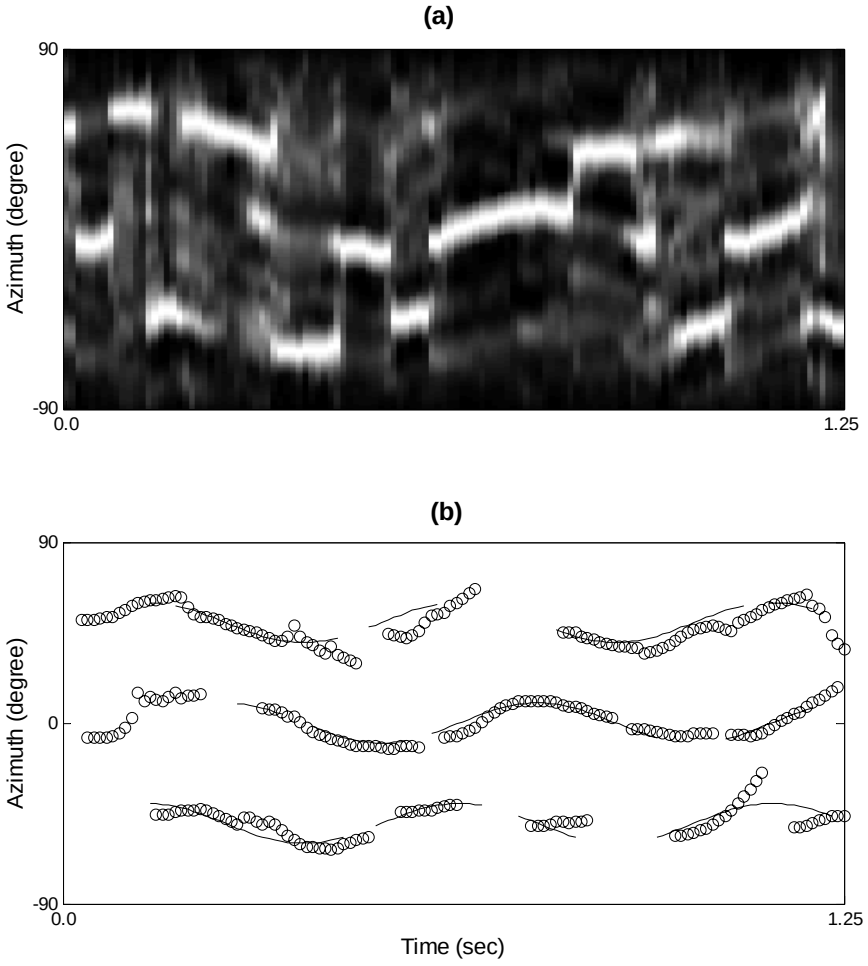

Fig. 10. Tracking three non-stationary sources using a Kalman filter approach. (a) Summarized cross-correlation across time. (b) Continuous tracks using the Kalman filter approach. The solid lines show the true trajectories where a gap indicates a pause in the sentence. The 'o' tracks correspond to the estimated source locations.

the cross-correlation function is replaced with a narrow-width Gaussian and all the individual channels are summed together. The results for the summary cross-correlation across time are shown in Fig. 10(a). Here the brighter regions correspond to stronger activities. For an anechoic situation, strong peaks are usually well correlated with the active sources. Hence, at each time frame we select all the azimuths corresponding to the prominent peaks in the summary cross-correlation function. As seen in Fig. 10(a), this representation exhibits spurious as well as missing peaks for a considerable number of frames. Smoothing these observations using Kalman filtering improves the location estimation. In Sturim et al., the Kalman filter is used for the tracking of single source tracks [15]. Specifically, we use a second-order auto-regressive model for the source motion. In addition, a data association algorithm is used to initialize and terminate tracks. The new observations are associated with individual tracks using acceptance regions that take into account the variance of measurement noise and the possible target motion [15].

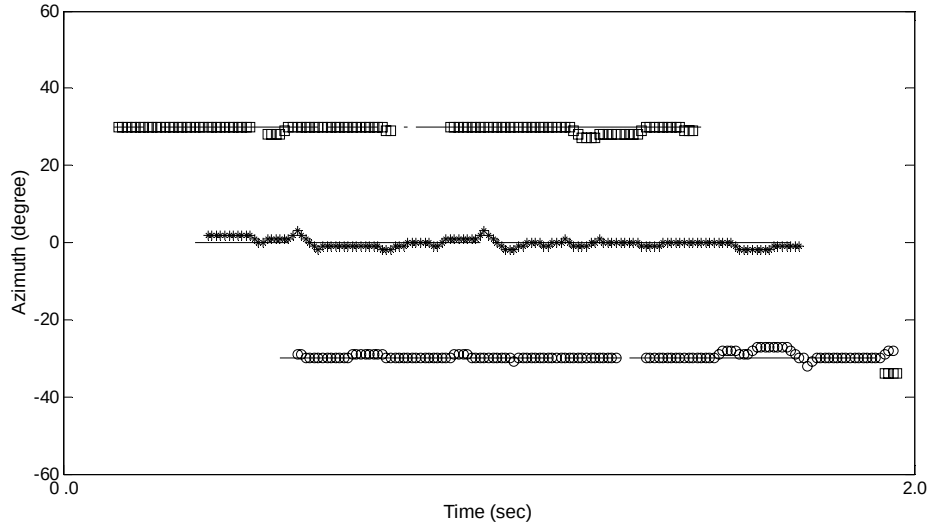

Fig. 11. Source tracking for three stationary sources. The solid lines show the true trajectories where a gap indicates a pause in the sentence. The '∗', 'o' and '□' tracks correspond to the estimated tracks.

Observations that cannot be associated with any of the active tracks are used in the initialization of a new track. The estimated tracks obtained using this approach are presented in Fig. 10 (b).

Note that in the Kalman filter approach presented above there is no correspondence between estimated tracks across time. This differs from our system which uses the continuity of the tracks at the boundaries between the one-, two- and three-source subspaces to reconstruct the individual tracks across time. A comparison between Fig. 10(b) and Fig. 9(d) also shows that our HMM model performs substantially better in estimating the individual source locations.

VII. DISCUSSION

We have proposed a new approach for tracking multiple moving sound sources. Our approach includes an across-frequency statistical integration method for localization and an HMM framework that imposes continuity constraints across time for individual tracks along with a switching mechanism for transition between subspaces corresponding to different numbers of active sources. As a result, the system is able to automatically detect the number of active sources at a given time and estimate their locations. Such a property is highly desirable in speech applications where speakers spontaneously change locations and utter words in a sporadic way.

Our system may also be applied to the multi-source localization of stationary sources. Fig. 11 shows such an example with three stationary sources: one female speaker at -30°, one male speaker at 0°, and another female speaker at 30°. The signals for the three sources are equalized to have the same average energy at the two ears. To demonstrate the system capability to jump between the subspaces with zero, one, two and three sources, we let the three speech utterances start and end at different times. As shown in the figure, the system correctly detects the number

of sources for a majority of time frames. Moreover, the source locations are estimated to within 5° of true azimuths. This demonstrates the potential of our system in localizing stationary sources. A standard localization method for stationary sources summates the cross-correlations across both frequency and time [18]. Each prominent peak in the resulting summary cross-correlation indicates an active source. However, such pooling often leads to spurious or missing peaks, which in turn result in significant tracking errors. Tracking of individual sources across time as well as detection of the number of sources at a given time provides a more detailed description which may be necessary for improved accuracy.

While the current system does not consider reverberation, our framework holds promise for reverberant conditions. Under reverberation, ITD and IID cues become noisy due to the multiple reflections of a sound source. However, the acoustic onsets are generally unaffected by the reflections and thus could be utilized to trigger ITD and IID estimation during intervals where reverberant energy is weak. Therefore, an onset detector could be incorporated in our channel selection stage in order to improve the localization of reverberant sound sources.

Although we have considered a maximum of three sources, our tracking framework is extensible to an arbitrary number of sources. With increased number of sources, the number of reliable channels decreases and hence the dynamics part of the model should play a more dominant role. However, the state space grows exponentially with the number of sources and thus efficient pruning strategies will become increasingly necessary. Also, the system needs to incorporate additional information in order to robustly identify possible direction changes at crossing points, such as spectral and pitch continuity. These issues as well as tests on sound motions in real environments require further research.

ACKNOWLEDGMENTS

A preliminary version of this work was presented at ICASSP 2003. This research was supported in part by an AFOSR grant (FA9550-04-1-0117) and an NSF grant (IIS-0081058).

REFERENCES

[1] Omologo, M., Svaizer, P. and Matasoni, M., "Environmental conditions and acoustic transduction in hands-free speech recognition," *Speech Commun.*, vol. 25, no. 1-3, pp. 75-95, 1998.

[2] Ajmera, J., Lathoud, G. and McCowan, L., "Clustering and segmenting speakers and their locations in meetings," *Proc. ICASSP*, vol. 1, pp. 605-608, 2004.

[3] Stone, L. D., "A Bayesian approach to multiple-target tracking," in *Handbook of Multisensor Fusion*, Hall, D. L. and LLinas, J., Eds., Boca Raton, FL: CRC Press, 2001.

[4] Koch, W., "Target tracking," in *Advanced Signal Processing Handbook*, Stergiopoulos, S., Ed., Boca Raton, FL: CRC Press, 2001.

[5] Reid, D., "An algorithm for tracking multiple targets," *IEEE Trans. Automatic Control*, vol. 24, no. 6, pp. 84-90, 1979.

[6] Bar-Shalom, Y. and Tse, E., "Tracking in a cluttered environment with probabilistic data association," *Automatica*, vol. 11, pp. 451-460, 1975.

[7] Buckley, K., Vaddiraju, A. and Perry, R., "A new pruning/merging algorithm for MHT multitarget tracking," *Proc. IEEE International Radar Conference*, pp. 71-75, 2000.

[8] Gordon, N., "A hybrid bootstrap filter for target tracking clutter," *IEEE Trans. Aerosp. Electron. Syst.,* vol. 33, no.1, pp. 353-358, 1997.

[9] Isard, M. and Blake, A., "Condensation - conditional density propagation for visual tracking," *Int. J. Computer Vision*, vol. 29(1), pp.5-28, 1998.

[10] Martinerie, F., "Data fusion and tracking using HMMs in a distributed sensor network," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 33(1), pp. 11-28, 1997.

[11] Knapp, C. and Carter, G., "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 24(4), pp. 320- 327, 1976.

[12] Krim, H. and Viberg, M., "Two decades of array signal processing research: The parametric approach," *IEEE Signal Processing Mag.*, vol. 13, pp. 67-94, 1996.

[13] Vermaak, J. and Blake, A., "Nonlinear filtering for speaker tracking in noisy and reverberant environments," *Proc. ICASSP*, vol. 5, pp. 3021-3024, 2001.

[14] Ward, D. B., Lehmann, E. A. and Williamson, R. C., "Particle filtering algorithms for tracking and acoustic source in a reverberant environment," *IEEE Trans. Speech Audio Process*., vol. 11, no. 6, pp. 826-836, 2003.

[15] Sturim, D. E., Brandstein, M. S. and Silverman, H. F., "Tracking multiple talkers using microphone-array measurements," *Proc. ICASSP*, vol. 1, pp. 371-374, 1997.

[16] Potamitis, I., Tremoulis, G. and Fakotakis, N., "Multi-array fusion for beamforming and localization of moving speakers," *Proc. Eurospeech*, vol. 2, pp. 1721-1724, 2003.

[17] Nakadai, K., Hidai, K., Mizoguchi, H., Okuno, H. G. and Kitano, H., "Real-time auditory and visual multiple-object tracking for humanoids," *Proc. of 17th IJCAI*, pp. 1425-1432, 2001.

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

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

[20] Burkhard, M. D. and Sachs, R. M., ‘‘Anthropometric manikin for acoustic research,’’ *J. Acoust. Soc. Am.*, vol. 58, 214–222, 1975.

[21] Wu, M., Wang, D. L. and Brown, G. J., “A multipitch tracking algorithm for noisy speech,” *IEEE Trans. Speech Audio Process.*, vol. 11, pp. 229-241, 2003.

[22] Wu, M., “Pitch tracking and speech enhancement in noisy and reverberant environments,” Ph.D. dissertation, The Ohio State University, Computer and Information Science, 2003.

[23] Roweis, S. T., "One-microphone source separation," in *Advances in Neural Information Processing Systems 13 (NIPS’ 00)*, MIT Press , pp. 793—799, 2001.

[24] Gardner, W. G. and Martin, K. D., “HRTF measurements of a KEMAR dummy-head microphone,” *MIT Media Lab Perceptual Computing Technical Report* #280, 1994.

[25] Gilkey, R. H. and Anderson, T. R., Eds*., Binaural and Spatial Hearing in Real and Virtual Environments*, Mahwah, NJ: Lawrence Erlbaum Associates, 1997.

[26] Mehrgardt, S. and Mellert, V., “Transformation characteristics of the external human ear,” *J. Acoust. Soc. Am.*, vol. 61, pp. 1567-1576, 1977.

[27] Begault, D. R., *3-D Sound for Virtual Reality and Multimedia*, Academic Press, 1994.

[28] Kistler, D. J. and Wightman, F. L., “A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction,” *J. Acoust. Soc. Am.*, vol. 91, pp. 1637-1647, 1992.

[29] Gold, B. and Morgan, N., *Speech and Audio Signal Processing*, New York: John Wiley and Sons, 2000.

[30] Moore, B. C. J., *An introduction to the Psychology of Hearing*, 5th ed., San Diego, CA: Academic, 2003.

[31] Patterson, R. D., Nimmo-Smith, I., Holdsworth, J. and Rice, P., ”An efficient auditory filterbank based on the gammatone function,” *APU Report* 2341, Cambridge: Applied Psychology Unit, 1988.

[32] Moore, B. C. J., Glasberg, B. R. and Baer, T., “A model for prediction of thresholds, loudness and partial loudness,” *J. Audio Eng. Soc*., vol. 45, pp 224-240, 1997.

[33] Rabiner, L. R. and Juang, B. H., *Fundamentals of Speech Recognition*, 2nd ed., Englewood Cliffs, NJ: Prentice-Hall, 1993.

[34] Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D. and Dahlgren, N., "Darpa timit acoustic-phonetic continuous speech corpus," *Technical Report NISTIR* 4930, National Institute of Standards and Technology, Gaithersburg, MD, 1993.

[35] MacPherson, E. A., "A computer model of binaural localization for stereo imaging measurement," *J. Audio Eng. Soc.*, vol. 39, pp. 604-622, 1991.

[36] Hand, D. J. and Yu K., "Idiot's Bayes – not so stupid after all?" *Int. Stat. Review*, vol. 69, pp. 385-398, 2001.

[37] Jelinek, F., *Statistical Methods for Speech Recognition*, MIT Press, 1997.

[38] Russel, S. J. and Norvig, P., *Artificial Intelligence: A Modern Approach*, 2nd ed., Prentice Hall, 2002.