Binaural Tracking of Multiple Moving Sources

Nicoleta Roman and DeLiang Wang, *Fellow, IEEE*

***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 all 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. Results are presented for up to three moving talkers in anechoic conditions. 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, multisource tracking.**

I. Introduction

THE problem of tracking multiple moving targets arises in many domains including surveillance, navigation, and speech processing. In this paper, 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 multimodal distributions is provided using sequential Monte Carlo (SMC) 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. Recently, a Bayesian tracking framework based on the random finite set (RFS) theory has been proposed to statistically model the time-varying number of sources [10]. In the RFS approach, the multitarget states and the multitarget measurements are modeled as random finite sets that change their cardinality as well as their values with time. As in standard Bayesian filtering, SMC implementations of the Bayesian RFS filter produce computationally tractable solutions. However, as the number of objects increases the Bayesian RFS filter becomes expensive to implement. A more tractable alternative is the probability hypothesis density (PHD) filter which propagates only the first moment of the multitarget posterior [10].

HMM has been proposed for multiple frequency line tracking as well as target tracking in sonar networks by employing the Markovian modeling of source dynamics in a discretized target space [11], [12]. It is important to note that this framework can handle multimodal likelihood distributions. Due to discrete Markov modeling, Viterbi decoding can be used to efficiently search for the most likely state sequences. Track initiation and termination can be modeled probabilistically in an HMM framework by, for example, introducing transitions from and to a terminating state [11].

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 [13] or subspace-based methods [14]. Algorithms that combine Kalman filtering with probabilistic data association techniques have been proposed in [15] and [16] for the tracking of multiple speakers. The particle filtering theory has been applied to the tracking of one moving speaker in a reverberant environment (see [17], [18]) as well as to the tracking of an unknown number of moving speakers [19]. The RFS Bayesian tracker has also been applied to the problem of tracking an unknown time-varying number of speakers [20]. These multisource tracking algorithms have

Manuscript received March 26, 2007; revised December 19, 2007. This work was supported in part by the Air Force Office of Scientific Research under Grant FA9550-04-1-0117 and the National Science Foundation under Grant IIS-0081058. A preliminary version of this work was presented at ICASSP 2003. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Israel Cohen.

N. Roman is with the Department of Mathematics, The Ohio State University, Lima, OH 45804 USA (e-mail: roman.45@osu.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 USA (e-mail: dwang@cse.ohio-state.edu).

Digital Object Identifier 10.1109/TASL.2008.918978

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 multisource tracking problem becomes more challenging and few results have been published in this area. A recent study [21] utilizes multidimensional statistical filtering methods to simultaneously track source directions and source spectral envelopes for up to two concurrent speakers from binaural input. Solutions that combine both visual and auditory information, where audition helps mainly in resolving ambiguities during occlusion, have also been proposed [22].

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 [23], [24]. 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 [25]. 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 most 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 multipitch tracking proposed by Wu *et al.* [26], [27]. This approach resembles the HMM tracking system proposed in [11]. The system in [11] is, however, tracking a mixed track whereas our system aims at producing continuous tracks for each active source separately. Due to the sparsity of speech signal distribution in a 2-D T–F representation [28], some T–F units in a mixture signal 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 overlapping time sections 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, for a given time frame, 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. A statistical approach to binaural cue integration for sound localization has been proposed in [29]. Due to subband analysis, such binaural approaches have strong potential for multisource localization. By comparison, standard beamforming-based delay-of-arrival estimators such as variations of the MUSIC and ESPRIT algorithms require that the number of sources is fewer than the number of microphones.

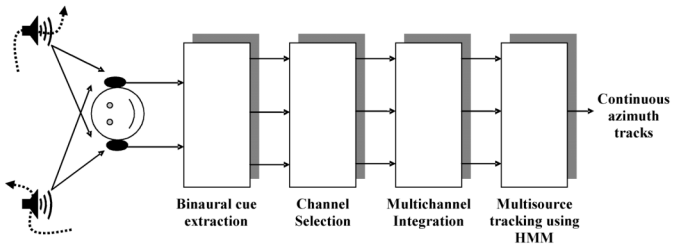

Fig. 1. Schematic diagram of the proposed multisource tracking system.

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. Section VI gives results for tracking up to three simultaneous speakers in various conditions as well as a comparison with a Kalman filter approach. The last section concludes the paper.

II. Model Architecture

Our multisource 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 multisource 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 [30] 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 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 $[-90^\circ, 90^\circ]$. 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 (see, e.g., [29]). Hence, for each moving source, left and right ear signals are obtained by filtering with time-varying HRTFs that correspond to the 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, an auditory periphery model is used to obtain a frequency decomposition of the left and right ear mixtures. 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 higher frequencies, multiple peaks are present and this creates ambiguity in localization. In addition, IID is computed using the energy ratio between the two ears independently in each T–F unit.

Frequency channel selection comprises the second stage of our system. This stage attempts to select reliable channels de-

fined 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 [23]. 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 observation 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 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 [31]. 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 [30] 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° azimuth. The sampling rate is 44.1 kHz.

An attractive property of HRTFs is that they are almost minimum-phase [32]. 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 [33]. 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 [34]. 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 [35]. 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 filter is changed every time sample. 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 [36]. The filterbank employed here consists of 128 fourth-order gammatone filters [37] 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 [38]. 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. We note that, while the auditory system appears to use the interaural differences between response envelopes at high frequencies, the system proposed extracts ITD and IID cues from the responses directly, hence sensitive to the fine structure of the responses.

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, as shown by (1) at the bottom of the page, where l_c, r_c refer to the left and right peripheral output for channel c, and $\bar{l}_c, \bar{r}_c$ 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. Peaks are identified by comparing a value with its two neighboring values in the cross-correlation function, corresponding to a derivative operation. At higher frequencies, localization of narrowband sources is intrinsically ambiguous due

$$C(c,m,\tau)=\frac{\sum_{k=0}^{K-1}\left(l_c(m\cdot K/2-k)-\bar{l}_c\right)\left(r_c(m\cdot K/2-k-\tau)-\bar{r}_c\right)}{\sqrt{\sum_{k=0}^{K-1}\left(l_c(m\cdot K/2-k)-\bar{l}_c\right)^2}\sqrt{\sum_{k=0}^{K-1}\left(r_c(m\cdot K/2-k-\tau)-\bar{r}_c\right)^2}} \tag{1}$$

to the periodic nature of the cross-correlation function. 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

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

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 [39]. 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. 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.

A. Dynamics Model

In a practical multisource 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:

$$S = S_0 \cup S_1^1 \cup S_1^2 \cup S_1^3 \cup S_2^{1,2} \cup S_2^{1,3} \cup S_2^{2,3} \cup S_3 \quad (3)$$

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 ith 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. Note that s_m is one of the eight possible subspaces described above. Then the state transitions are described by

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

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

TABLE I
Jump Probabilities Between Subspaces With Zero, One, Two, and Three Active Sources

	$\to S_0$	$\to S_1^1$	$\to S_1^2$	$\to S_1^3$	$\to S_2^{1,2}$	$\to S_2^{1,3}$	$\to S_2^{2,3}$	$\to S_3$
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 30 mixtures; each mixture consists of three speech utterances from the TIMIT database with an average duration of 2.5 s [40]. The speech utterances are selected to have similar lengths in order to maximize the overlap. 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 ith source is described by

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

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

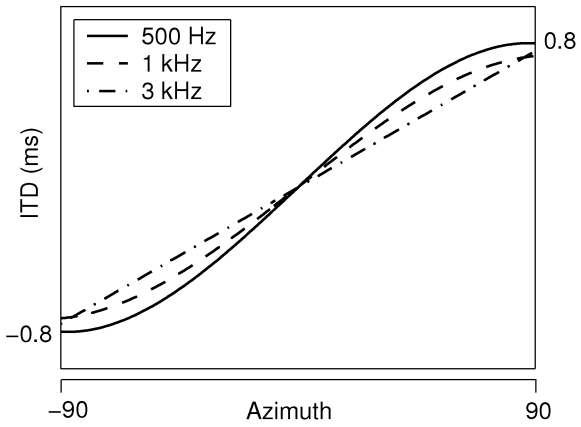

Fig. 2. ITD reference functions for three auditory channels with center frequencies of 500 Hz, 1 kHz, and 3 kHz and azimuth in the range $[-90°, 90°]$.

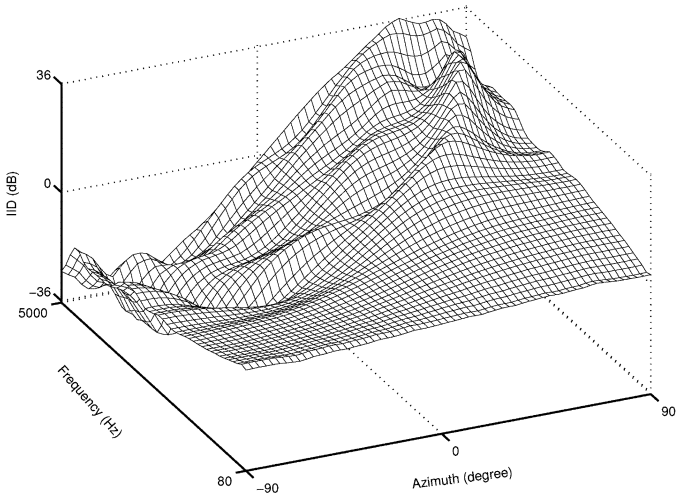

Fig. 3. IID reference functions for frequency in the range 80 Hz–5000 Hz and azimuth in the range $[-90°, 90°]$.

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. Moreover, ITD and IID deviate from the actual ITD/IID reference values and can indicate phantom sources [23]. 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. The threshold varies between 0.95 and 0.5.

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. 2 shows ITD values for three auditory channels with center frequencies 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 [41]. IID reference values for all frequency channels are also shown in Fig. 3. 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 $\tau_{\mathrm{ref}}(c,\varphi)$ and $\iota_{\mathrm{ref}}(c,\varphi)$. Given a set of selected peaks and the estimated ITD in a T–F unit, we define the ITD and IID deviations as

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

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

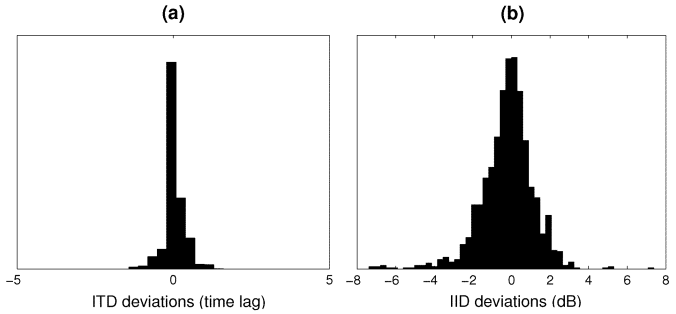

Fig. 4. Histogram of estimated ITD and IID deviations from reference values for a channel with $f_c = 1.5$ kHz in the one-source scenario.

where τ is the lag of the closest peak in the cross-correlation function to $\tau_{\mathrm{ref}}(c,\varphi)$ and ι is the estimated IID. Statistics of the deviations δ_τ and δ_ι are collected separately for each frequency channel across different time frames. Fig. 4 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 ten 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 $\Delta_\tau = \max((f_s)/(2f_c),44)$, 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

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

TABLE II
ESTIMATED MODEL PARAMETERS FOR ONE-SOURCE AND MULTISOURCE 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

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 multisource scenarios with two and three active sources. We employ a set of ten 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 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 multisource scenarios assuming a fixed noise level q across all conditions and all frequency channels. This ensures that the background noise and the unreliable channels do not influence the comparison between one-source and multisource scenarios. ML estimation is implemented empirically using the distribution given in (7) for the ITD/IID deviations in the datasets described above. This estimation gives $q = 0.03$. The parameters a_1, a_2, a_3, and a_4 are reported in Table II.

C. Likelihood Model

In this subsection, we derive the conditional probability density $p(\{T_c, \iota_c\} \,|\, \boldsymbol{x})$, often referred to as the likelihood, which statistically describes how 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 $p(\{T_c, \iota_c\} \,|\, \boldsymbol{x})$ for the one-source subspaces, i.e., $\boldsymbol{x} \in S_1^1 \cup S_1^2 \cup S_1^3$. For channel c, we compute the deviations $\delta_\tau, \delta_\iota$ as described in (6) using as reference values $\tau_{\text{ref}}(c, \varphi)$ and $\iota_{\text{ref}}(c, \varphi)$, 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

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

where the symbols are as described in (7) and (9) and the parameters are estimated for the one-source scenario. Note that the uniform background noise is assigned to an unreliable channel in order to weigh similarly the noisy and the nonselected channels.

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 [42]:

$$p(\{T_c, \iota_c\} \,|\, \boldsymbol{x}) = \kappa^{N_b} \sqrt{\prod_c p(T_c, \iota_c \,|\, \boldsymbol{x})} \tag{11}$$

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

Next, we consider the conditional probability $p(\{T_c, \iota_c\} \,|\, \boldsymbol{x})$ for the two-source case, i.e., $\boldsymbol{x} \in S_2^{1,2} \cup S_2^{1,3} \cup S_2^{2,3}$. Similar to the one-source case, we compute the deviations δ_τ^k and δ_ι^k with respect to the kth 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 kth 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 source. Specifically, we label channel c as belonging to the kth source if the corresponding deviations satisfy $|\delta_\tau^k| < \varepsilon\lambda_\tau(c)$ and $|\delta_\iota^k| < \varepsilon\lambda_\iota(c)$, where $\varepsilon = 5$ is the gate size. Assume that the kth source is the stronger among the two (most selected channels are dominated by the kth source). Then the conditional probability for channel c under this assumption is given by

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

where all the parameters are derived for the multisource case. Consider the case of a single active source at azimuth angle Φ_1 such that all the estimates show only small deviations from the reference values. The gating ensures that for a hypothesized two-source state that contains azimuth Φ_1, all the estimates are associated with the Φ_1 source. In this case, the conditional probability for a selected channel is evaluated using the second line of (12) and thus is computed only based on deviations from Φ_1 irrespective of the value of the second azimuth. We thus avoid fitting the data with a model of two-closely spaced sources when only one source is active.

We apply integration of the individual probabilities across all channels as done in (11) to give the conditional probability

$p(\{T_c, \iota_c\} \,|\, \boldsymbol{x}, k)$ for the current time frame under the assumption that the kth hypothesized source is the strongest. Finally, the conditional probability $p(\{T_c, \iota_c\} \,|\, \boldsymbol{x})$ for the current time frame is the larger of assuming either the first or the second hypothesized source to be the stronger source

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

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

Note that, without the gating mechanism, (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. As seen above, the gating mechanism is one way to penalize the overfitting due to noise.

Similar to the two-source case, we consider the conditional probability $p(\{T_c, \iota_c\} \,|\, \boldsymbol{x})$ for the three-source case, i.e., $\boldsymbol{x} \in S_3$. Equations (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 set 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., $x \in S_0$

$$p(\{T_c, \iota_c\} \,|\, \boldsymbol{x}) = \kappa\alpha_0 \tag{14}$$

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 using 1° spacing for azimuth and 10-ms time frames and the standard Viterbi algorithm is employed in order to identify the optimal sequence of states [43]. 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 Viterbi algorithm used here is a batch algorithm, although online versions exist that trade the precision of the solution with speed [44].

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 $[-3\sigma, 3\sigma]$,

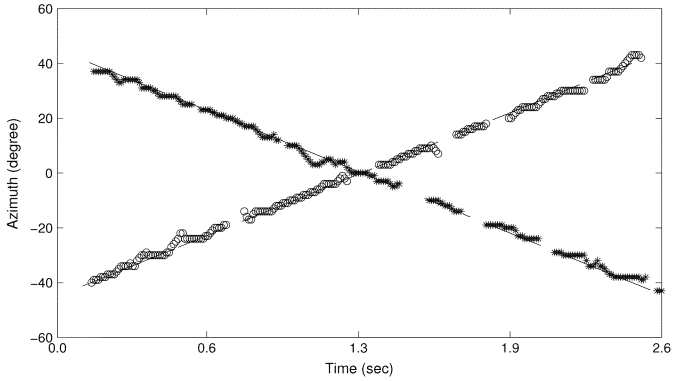

Fig. 5. Source tracking for two crossing sources with linear motion. The solid lines show the true trajectories where a gap indicates a pause in the sentence. The "$*$" and "o" tracks correspond to the estimated tracks.

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 [45]. 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 one-, two- and three-source scenarios. As described in Section III, binaural synthesis is used to simulate 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. 5 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 and anechoic conditions are assumed. 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 to track the sources across time as long as they are not 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 frequency channels by the other source. While in Fig. 5 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 are used during training, our system works for nonlinear motions. Fig. 6 shows the result of tracking one female and one male speaker moving on nonlinear trajectories consisting of two cosine azimuth paths 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 intersection point. However, as seen in Fig. 5, our system could disambiguate between two tracks at a crossing

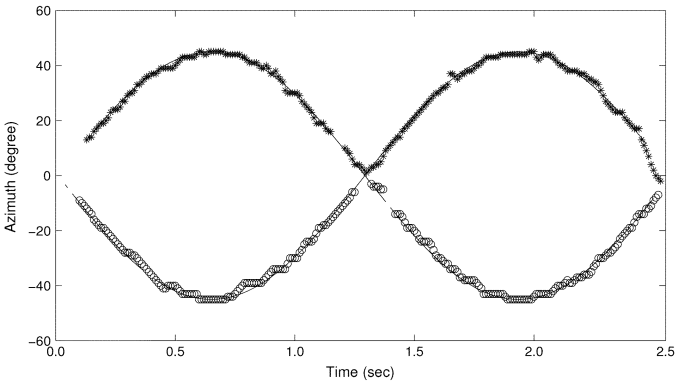

Fig. 6. Source tracking for two crossing sources with nonlinear motion. The solid lines show the true trajectories where a gap indicates a pause in the sentence. The "∗" and 'o" tracks correspond to the estimated tracks.

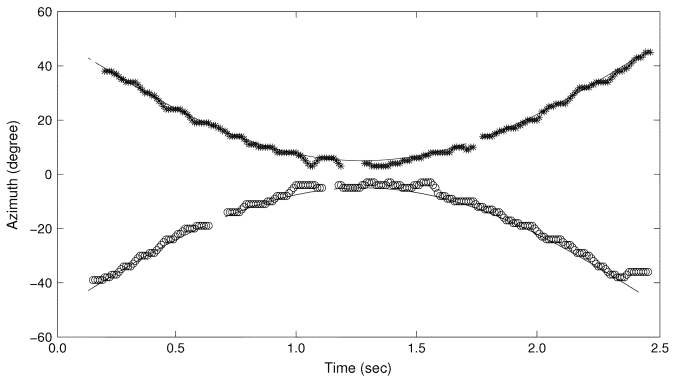

Fig. 7. Source tracking for two sources with closely spaced motions. The solid lines show the true trajectories where a gap indicates a pause in the sentence. The "∗" and "o" tracks correspond to the estimated tracks.

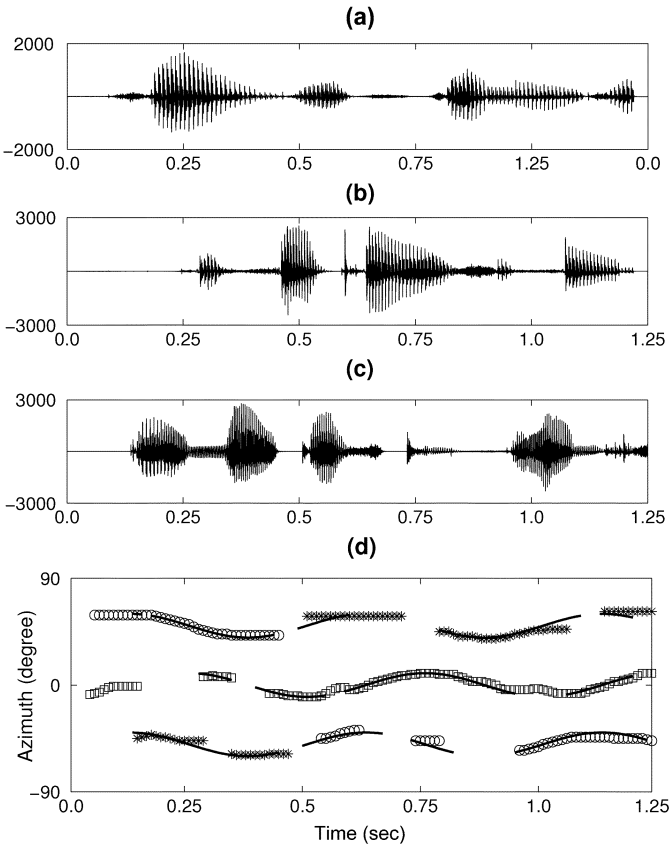

Fig. 8. Tracking three nonstationary moving sources. (a) Left ear signal for the first speaker. (b) Left ear signal for the second speaker. (c) Left ear signal for the third speaker. (d) Continuous tracks obtained by the proposed model. The solid lines show the true trajectories where a gap indicates a pause in the sentence. The "∗," "o," and "□" tracks correspond to the estimated tracks.

point when the likelihood is dominated by a single continuous source in the neighborhood of the point. In Fig. 5, the source corresponding to the "o" 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. Additional features could be incorporated into the model to help correctly associate targets with tracks at the crossing points, such as spectral continuity of individual sources.

Fig. 7 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. 8 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. 8(a), (b) and (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 one-source subspace and the three-source subspace. In Fig. 8, 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. 8(d).

We have systematically evaluated the performance of the proposed system. Although our model deals with multisource situations explicitly, diffuse background is often present in acoustic environments. To examine how our system functions in such environments, our evaluation also includes some background noise. We report results for the following conditions: 1) one moving speaker with no background noise as well as with background noise at 40, 20, and 10 dB SNR, respectively; 2) two moving speakers with an interspeaker separation of 30° and no background noise as well as background noise of 40-dB SNR; 3) two moving speakers with an interspeaker separation of 10° and no background noise as well as background noise of 40-dB SNR; and 4) three moving speakers with an interspeaker separation of 30° and no background noise as well as background

TABLE III
SYSTEMATIC EVALUATION OF THE PROPOSED MODEL

	Overestimation (%)	Underestimation (%)	Accuracy (°)
1 speaker (no noise)	5%	2%	0.53
1 speaker (40-dB SNR)	3%	4%	0.67
1 speaker (20-dB SNR)	29%	7%	1.24
1 speaker (10-dB SNR)	37%	23%	2.62
2 speakers (30° separation, no noise)	5%	13%	2.47
2 speakers (30° separation, 40-dB SNR)	7%	13%	2.94
2 speakers (10° separation, no noise)	4%	21%	1.63
2 speakers (10° separation, 40-dB SNR)	4%	21%	1.64
3 speakers (no noise)	3%	50%	7.18
3 speakers (40-dB SNR)	3%	55%	6.92

noise of 40-dB SNR. The background noise is formed by adding white noise independently to each of the two ear signals at a specified SNR level. We assume that the speakers are moving at a speed of 1 m/s on the frontal semicircle around the KEMAR and thus the tracks for all the conditions here are linear; for example, in the one-speaker scenario, the trajectory goes from 0° to 90° azimuth. A total of ten mixtures are used for each condition, and the results are presented in Table III. The binaural signals are simulated as before using speech utterances from the TIMIT database. The left ear signals for individual sources are equalized before summation. Table III shows the percentage of frames where the number of sources is overestimated or underestimated, separately. Additionally, the accuracy of estimating the source locations is given in the last column. At each frame, the estimated locations are compared against the actual locations and the standard deviation is reported in the column. When the number of sources is incorrectly estimated, the estimated azimuths closest to the actual ones are used in the calculation of the standard deviation. The results in Table III show that, although the error of estimating the number of sources increases going from one source to three sources, the accuracy of tracking remains reasonable. In addition, the system is robust to moderate levels of background noise.

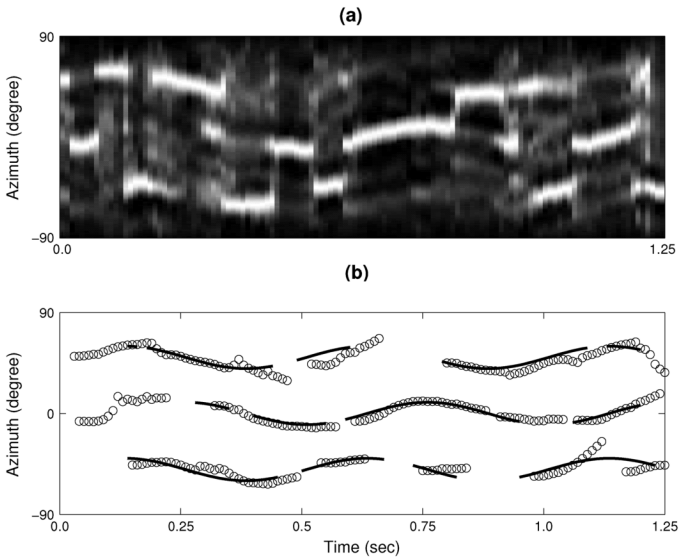

Fig. 9. Tracking three nonstationary sources using a Kalman filter approach. (a) Summarized cross correlation across time. (b) Continuous tracks using the Kalman filter approach. The solid lines show the true trajectories where a gap indicates a pause in the sentence. The "o" tracks correspond to the estimated source locations.

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. 9 shows the extracted tracks using this Kalman filtering approach for the same three source configuration as used in Fig. 8. For azimuth estimation, we employ the skeleton cross-correlogram described in [23] 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 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. 9(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. 9(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 autoregressive 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]. 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. 9(b).

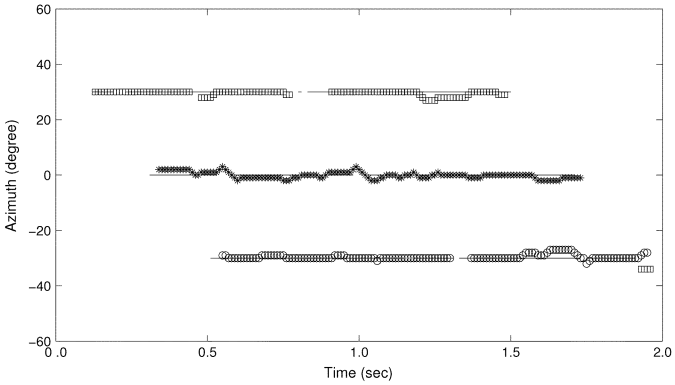

Fig. 10. Source tracking for three stationary sources. The solid lines show the true trajectories where a gap indicates a pause in the sentence. The "$*$," "o," and "□" tracks correspond to the estimated tracks.

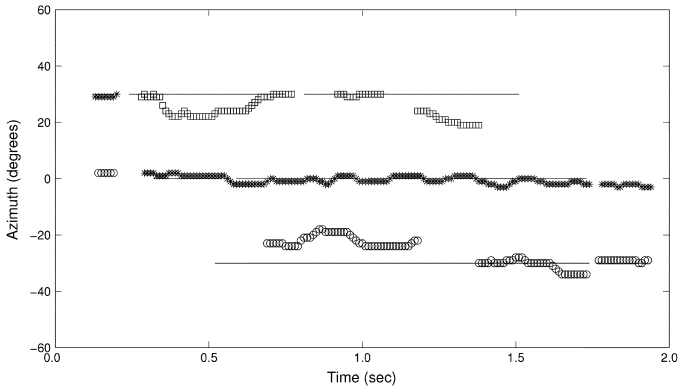

Fig. 11. Source tracking for three stationary sources in a reverberant condition ($\mathrm{T}_{60} = 50$ ms). The solid lines show the true trajectories where a gap indicates a pause in the sentence. The "$*$," "o," and "□" tracks correspond to the estimated tracks.

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 Figs. 9(b) and 8(d) shows that our HMM model performs better in estimating the individual source locations. An evaluation of the Kalman filter approach on the three-speaker tracking task described in Table III results in 17% overestimation, 31% underestimation, and an accuracy of 12.87° with no background noise. With 40-dB SNR background noise, the Kalman filter yields 17% overestimation, 32% underestimation, and an accuracy of 10.14°. The slightly higher accuracy with background noise is probably due to sensitivity to the track initialization and termination rules. Our model produces more accurate tracking, while its underestimation error is substantially higher and overestimation error is significantly lower.

VII. Conclusion

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 provide accurate location estimates. 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 multisource localization of stationary sources. Fig. 10 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 the 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 [23]. 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 detecting the number of sources in each time frame gives a detailed description which may be necessary for improved accuracy in multitalker scenarios.

The current system does not consider reverberation. To examine whether the system can tolerate mild reverberation, we simulate binaural mixtures for the same three source scenario described above in a mildly reverberant condition. The left and right reverberated impulse responses are simulated using the image acoustic model described in [46] on the same HRTF database used above. The reverberation time is $\mathrm{T}_{60} = 50$ ms (T_{60} is the time required for the sound level to drop by 60 dB following the sound offset). The results are given in Fig. 11, and show that our system is able to estimate the number of sources in most frames at the expense of decreased accuracy. Under reverberation, ITD and IID cues are smeared due to the multiple reflections of a sound source [47]. Our channel selection based on cross correlation can be used as a simple technique for integrating the binaural cues in T–F units with relatively little reverberation. Consequently, as seen in Fig. 11, our system is able to function to some extent. However, future research is required in order to make the system robust to room reverberation. It is well known that the acoustic onsets are generally unaffected by reflections, and hence they can inform the system on when to perform localization so as to minimize the adverse effects of reverberation. In [48], robust acoustic onset detection is investigated for binaural localization which incorporates a physiologically inspired model for a precedence effect. The coupling of such onset detectors with our statistical integration framework can potentially improve the performance of our system in reverberant conditions.

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 adaptation of our system to robust tracking in real environments require further research.

REFERENCES

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

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

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

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

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

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

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

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

[9] M. Isard and A. Blake, "Condensation—Conditional density propagation for visual tracking," *Int. J. Comput. Vision*, vol. 29, no. 1, pp. 5–28, Jan. 1998.

[10] B.-N. Vo, S. Singh, and A. Doucet, "Sequential Monte Carlo methods for Bayesian multi-target filtering with random finite sets," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 41, no. 4, pp. 1224–1245, Oct. 2005.

[11] X. Xie and R. J. Evans, "Multiple target tracking and multiple frequency line tracking using hidden Markov models," *IEEE Trans. Signal Process.*, vol. 39, no. 12, pp. 2659–2676, Dec. 1991.

[12] F. Martinerie, "Data fusion and tracking using HMMs in a distributed sensor network," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 33, no. 1, pp. 11–28, 1997.

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

[14] H. Krim and M. Viberg, "Two decades of array signal processing research: The parametric approach," *IEEE Signal Process. Mag.*, vol. 13, no. 4, pp. 67–94, Jul. 1996.

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

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

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

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

[19] J.-R. Larocque, J. P. Reilly, and W. Ng, "Particle filters for tracking an unknown number of sources," *IEEE Trans. Signal Process.*, vol. 50, no. 12, pp. 2926–2937, Dec. 2002.

[20] W.-K. Ma, B.-N. Vo, S. S. Singh, and A. Baddeley, "Tracking an unknown time-varying number of speakers using TDOA measurements: A random finite set approach," *IEEE Trans. Signal Process.*, vol. 54, no. 9, pp. 3291–3304, Sep. 2006.

[21] 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.

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

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

[24] A. S. Feng and D. L. Jones, , D. L. Wang and G. J. Brown, Eds., "Localization-based grouping," in *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. New York: Wiley/IEEE Press, 2006, pp. 187–207.

[25] M. D. Burkhard and R. M. Sachs, "Anthropometric manikin for acoustic research," *J. Acoust. Soc. Amer.*, vol. 58, pp. 214–222, 1975.

[26] 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.

[27] M. Wu, "Pitch tracking and speech enhancement in noisy and reverberant environments," Ph.D. dissertation, Comput. Inf. Sci., The Ohio State Univ., Columbus, 2003.

[28] S. T. Roweis, "One-microphone source separation," in *Advances in Neural Information Processing Systems 13 (NIPS'00)*. Cambridge, MA: MIT Press, 2001, pp. 793–799.

[29] J. Nix and V. Hohmann, "Sound source localization in real sound fields based on empirical statistics of interaural parameters," *J. Acoust. Soc. Amer.*, vol. 119, pp. 463–479, 2006.

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

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

[32] S. Mehrgardt and V. Mellert, "Transformation characteristics of the external human ear," *J. Acoust. Soc. Amer.*, vol. 61, pp. 1567–1576, 1977.

[33] D. R. Begault, Ed., *3-D Sound for Virtual Reality and Multimedia*. New York: Academic, 1994.

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

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

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

[37] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," Cambridge, APU Rep. 2341, 1988, Appl. Psychol. Unit.

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

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

[40] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren, "Darpa Timit Acoustic-Phonetic Continuous Speech Corpus," National Institute of Standards and Technology, Gaithersburg, MD, Tech. Rep. NISTIR 4930, 1993.

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

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

[43] F. Jelinek, *Statistical Methods for Speech Recognition*. Cambridge, MA: MIT Press, 1997.

[44] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithms, and System Development*. Upper Saddle River, NJ: Prentice-Hall PTR, 2001.

[45] S. J. Russel and P. Norvig, *Artificial Intelligence: A Modern Approach*, 2nd ed. Upper Saddle River, NJ: Prentice-Hall, 2002.

[46] K. J. Palomäki, G. J. Brown, and D. L. Wang, "A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation," *Speech Comm.*, vol. 43, pp. 361–378, 2004.

[47] C. Faller and J. Merimaa, "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *J. Acoust. Soc. Amer.*, vol. 116, pp. 3075–3089, 2004.

[48] B. Supper, T. Brookes, and F. Rumsey, "An auditory onset detection algorithm for improved automatic source localization," *IEEE Trans. Speech Audio Process.*, vol. 14, no. 3, pp. 1008–1017, May 2006.

Nicoleta Roman received the B.S. and M.S. degrees in computer science from the University of Bucharest, Bucharest, Romania, in 1996 and 1997, respectively, and the Ph.D. degree in computer science and engineering from The Ohio State University, Columbus, in 2005.

Since 2005, she has been with the Department of Mathematics, The Ohio State University, Lima. Her research interests include computational auditory scene analysis, binaural processing, and machine learning.

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

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

Dr. Wang received the U.S. National Science Foundation Research Initiation Award in 1992 and the U.S. Office of Naval Research Young Investigator Award in 1996. He also received the 2005 Outstanding Paper Award from the IEEE TRANSACTIONS ON NEURAL NETWORKS.