BINAURAL TRACKING OF MULTIPLE MOVING SOURCES

Nicoleta Roman and *DeLiang Wang*
Department of Computer and Information
Science and Center for Cognitive Science
The Ohio State University
Columbus, OH 43210, USA
{niki,dwang}@cis.ohio-state.edu

ABSTRACT

This paper presents a novel method for tracking the azimuth locations of multiple active sources based on binaural processing. Binaural cues are strongly correlated with source locations for spectral regions dominated by only one source. Therefore, this approach integrates reliable information across different frequency channels to produce a likelihood function in the target space. Finally, a hidden Markov model (HMM) is employed for forming continuous tracks and detecting the number of active sources across time. Experimental results are presented for simulated multi-source scenarios.

1. INTRODUCTION

Computational auditory scene analysis aims at separation of multiple sound streams, such as speech, environmental noise, music, etc. The auditory system is able to segregate the target signal from the acoustic mixture using various cues, including pitch, onset time and location. Experiments with stationary sources under anechoic conditions show good separation results when accurate locations are given [1]. However, in a realistic environment source motion and head movement must be considered. The goal of this paper is to examine the use of acoustical information for multi-source tracking.

Among the numerous tracking systems proposed, arrays of microphones have been shown to provide accurate results for locating sound sources in a number of scenarios [2][3]. When restricting the size of the array to only two sensors as humans have, the multi-source tracking problem becomes a challenging task and little has been attained in this respect. As a solution, fused visual and auditory information is generally used in this type of applications where the auditory stream helps mainly in resolving ambiguities during occlusions [4].

We present a novel auditory tracking system based on binaural cues extracted from responses of a KEMAR dummy head that realistically simulates the filtering process of the head, torso and external ear. A typical approach for tracking applications is Bayesian inference [5]. We study a model that has been successfully applied to multi-pitch tracking under noisy conditions [6].

The rest of the paper is organized as follows: the next section describes auditory motion modeling. Section 3 contains the model description and provides details of the statistical model employed. Section 4 gives simulation results and the last section concludes this paper.

2. MODELING AUDITORY MOTION

For human audition, sound source localization is primarily achieved with the binaural cues, of interaural time differences (ITD) and interaural intensity differences (IID). In this paper, source localization refers to azimuth localization in the horizontal plane. Measurements and models of head-related transfer functions (HRTF) are the standard method for a realistic binaural synthesis. We utilize here a catalogue of HRTF measurements collected by Gardner and Martin from a KEMAR dummy head under anechoic conditions [7].

For a moving sound, there are changes in ITD and IID that give velocity cues and enable the listener to perceive and track the source location. In addition, as the relative distance between the listener and the sound event changes there is a shift in frequency called the Doppler effect. For motions induced by human walking, however, the Doppler shift is negligible and thus not used in this study.

2.1 Binaural simulation

The transmission path for an HRTF measurement contains many subsystems, i.e. the loudspeaker, the microphone and the ear canal, that need to be compensated in order to obtain the desired response. Therefore, we use diffuse-field equalized HRTFs that eliminate all the factors that are not location-dependent.

An attractive property of the HRTFs is that they are almost minimum-phase [8]. 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 that implements the ITD [9]. 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 [10]. Here, the minimum-phase part is computed using the inverse Hilbert transform and the ITD is estimated as the mean of the excess phase in the range of interest from 80 Hz to 5 kHz.

The input to the system is monaural recordings sampled at 44.1 kHz. Binaural responses at the left and right eardrums are synthesized by filtering the signals with direction-dependent impulse responses $h_{L,R}(n;\varphi)$, where φ is a time-varying azimuth. The database of HRTF measurements has a resolution of 5^o. Separate tables for the corresponding minimum-phase filters and time delays are computed and a simple two-way linear interpolation is applied. For an arbitrary direction of sound

Proc. ICASSP2003, Volume 5, Pages 149-152, April 6-10, Hong Kong

incidence, the impulse response is reconstructed from the cascade of those two processes.

2.2 Auditory periphery and binaural processing

It is widely acknowledged that cochlear filtering can be modeled by a bandpass filterbank. The filterbank employed here consists of 128 fourth-order gammatone filters [11] with channel center frequencies (CCF) equally distributed on the 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 [12]. In the final step of the peripheral model, a simple model of hair cell transduction consists of half-wave rectification and a square root operation.

For each frequency channel *c*, a normalized cross-correlation function is computed in the plausible range of ITD from –1 ms to 1 ms. The lag τ_c ($-44 < \tau_c < 44$) of a peak in the cross-correlation function is a candidate for ITD estimation. At high frequencies where multiple peaks are present the set of all possible time lags $\{\tau_c\}$ is considered, and this creates ambiguity in localization. We resolve this ambiguity by using IID information, which is given by the ratio of signal power at the two ears. We use a window size of 20 ms for all computations and a 10 ms overlap between adjacent frames.

3. STATISTICAL TRACKING

Our statistical model for tracking multiple acoustic sources includes selecting reliable channels, a statistical model for the measurement error, a source dynamics model and a method of computing the likelihood of a particular time frame from the observed data. The last stage of the algorithm uses an HMM to form continuous tracks for all active sources present in the scene.

The height of the peak in the normalized cross-correlation function systematically decreases with the increase of noise level and thus represents a measure of reliability. Therefore, a channel *c* is considered reliable and thus selected if the corresponding peak height exceeds a threshold $\theta(c)$. The thresholds $\theta(c)$ are estimated so that the majority of "clean" channels are selected, where a clean channel is the one dominated by only one source, i.e. the relative strength of the source with respect to the interference is greater than a threshold R=0.8. We observe that $\theta(c)$ is a linearly decreasing function with respect to the CCF.

3.1 Measurement model

For each selected frequency channel, the measured ITD and IID signal a specific source location. By studying the deviations of the measurements from the reference values, we can derive the probability of one selected channel supporting a location hypothesis.

Consider channel *c* and azimuth φ for which the ITD and IID reference values are $\tau_{ref}(c,\varphi)$ and $I_{ref}(c,\varphi)$. We denote $\delta_1 = \tau - \tau_{ref}(c,\varphi)$, where τ is the lag of the closest peak in the cross-correlation function with respect to the reference value and $\delta_2 = I - I_{ref}(c,\varphi)$, where *I* is the computed IID.

Statistics of the deviations δ_1 and δ_2 in a particular channel are collected for one-source and two-source scenarios across various spatial configurations. We observe that the histograms obtained are sharply centered at zero; therefore, we model the distribution of the error measurement in channel *c* as a combination of a Laplacian distribution and a uniform distribution (see [6]):

$$p_c(\delta_1,\delta_2) = (1-q)L(\delta_1;\lambda_1(c))L(\delta_2;\lambda_2(c)) + qU(\delta_1,\delta_2), \quad (1)$$

where $0 < q < 1$ is a weighting coefficient. $L(\delta;\lambda)$ is the laplacian distribution with variance λ and *U* is the uniform distribution in the plausible range of δ_1 ([$-\Delta,\Delta$] step lags) and δ_2 ([-20, 20] dB), where $\Delta = \max(\frac{F_s}{2CCF(c)}, 44)$.

The parameters $\lambda_i(c)$ and the weighting factor q are estimated independently for all frequency channels. We observe that the variance for δ_1 decreases abruptly across channels whereas the trend for δ_2 is slowly increasing. To obtain smooth parameters across channels we use the following model:

$$\lambda_1(c) = A_0 + A_1 / CCF(c), \quad (2a)$$

$$\lambda_2(c) = B_0 + B_1 * CCF(c), \quad (2b)$$

The maximum likelihood method is then used to estimate (A_i, B_i, q) for the one-source and the two-source scenarios.

3.2 Dynamics model

In a practical multi-source tracking situation, the number of sources currently active is generally unknown. In this work, we assume a maximum of three sources and define the state space consisting of eight subspaces as follows:

$$S = S_0 \cup (S_1^i)_{i=1,..,3} \cup (S_2^{i,j})_{i,j=1,..,3} \cup S_3, \quad (3)$$

where S_0 is the silent state, S_1^i is the set of states tracking the *i*'th source, $S_2^{i,j}$ is the set of states simultaneously tracking both the *i*'th and the *j*'th source and S_3 the set of states tracking all three sources. A state is represented as a 3-dimensional vector $\boldsymbol{x} = (\varphi_1, \varphi_2, \varphi_3)$, where each dimension φ_i gives the azimuth location or indicates that the source is silent.

Systems with Markovian transition probabilities provide a standard statistical framework for dealing with multiple dynamic models. Suppose that the state of the system $\boldsymbol{x}_k$ at time t_k is in the subspace S_k. Then the system is summarized by:

$$p(\boldsymbol{x}_k, S_k \mid \boldsymbol{x}_{k-1}, S_{k-1}) = p(S_k \mid S_{k-1}) p(\boldsymbol{x}_k \mid \boldsymbol{x}_{k-1}), \quad (4)$$

where $p(S_k \mid S_{k-1})$ are the transition probabilities between the state subspaces and $p(\boldsymbol{x}_k \mid \boldsymbol{x}_{k-1})$ gives the temporal evolution of the state vectors. Here, we assume that the objects move independently of each other following a linear gaussian model:

$$p(\varphi_k \mid \varphi_{k-1}) = N(\hat{\varphi}_k, \sigma), \quad (5)$$

where N is the Gaussian distribution with mean $\hat{\varphi}_k$ and variance $\sigma = 2^o$. For slow moving targets, the predicted location $\hat{\varphi}_k$ is obtained using a linear estimate backtracked from state $\boldsymbol{x}_k$ during a time period of maximum 150 ms.

3.3 Likelihood model

This section derives the conditional probability density $p(M, C \mid \boldsymbol{x})$, often referred to as the likelihood function, which statistically describes what a single frame of measurements can say about the joint state x of the objects to be tracked. Here, C is the set of selected channels and M is the corresponding set of ITD and IID measurements $M = (\{\tau_c\}, I_c)_{c \in C}$.

First, we consider the conditional probability for one active source, i.e. $\boldsymbol{x} \in (S_1^i)_{i=1,..,3}$, $p(M, C \mid \varphi)$ where φ refers to the hypothesized azimuth. For channel c, we compute the measurement errors δ_1, δ_2 from the reference values $\tau_{ref}(c, \varphi)$ and $I_{ref}(c, \varphi)$ as described previously. Then, the conditional probability of a single channel with respect to the state $\boldsymbol{x}$ is derived as follows [6]:

$$p(\{\tau_c\}, I_c \mid \varphi) = \begin{cases} p_c(\delta_1, \delta_2), & \text{if channel } c \text{ is selected} \\ qU(\delta_1, \delta_2), & \text{else} \end{cases}, \quad (6)$$

where all the symbols are as described in Eq. 2 and Eq. 3 and the parameters are estimated from one-source scenarios. Note that background noise is assigned to an unreliable channel.

By combining the evidence from all the reliable channels, we obtain the likelihood function for one single time frame. We observe that assuming independence between channels results in noisy distributions, and thus the conditional probability is expressed as follows [6]:

$$p(M, C \mid \varphi) = K \sqrt[b]{\prod_{c=1}^{N} p(\{\tau_c\}, I_c \mid \varphi)}, \quad (7)$$

where b=25 is a smoothing factor, N=128 is the total number of channels, and K is a normalization factor.

Next, we consider the likelihood for the multi-source hypothesis, i.e. $\boldsymbol{x} \in ((S_2^{i,j})_{i,j=1,..,3} \cup S_3)$, $p(M, C \mid (\varphi_k)_{k=1,..,n})$ where φ_k corresponds to the azimuth of the k'th source. Similar to the one-source case, we compute the measurement deviations δ_1^k and δ_2^k from the reference values $\tau_{ref}(c, \varphi_k)$ and $I_{ref}(c, \varphi_k)$ for all n active sources $(n = 2,3)$. Observe that a selected channel should signal only one source under the assumption that only one speaker dominates a reliable channel.

Let φ_1 be the strongest source in the current time frame. A gating technique is applied to label the channels that belong to the strongest source. Specifically, we label channel c if both $|\delta_1^1| < \beta \lambda_1(c)$ and $|\delta_2^1| < \beta \lambda_2(c)$ where $\beta = 5$ is the gate size. Then the conditional probability for channel c, when assuming φ_1 is stronger, is given by [6]:

$$p'(\{\tau_c\}, I_c \mid (\varphi_k)_k) = \begin{cases} qU(\delta_1^1, \delta_2^1), & \text{if channel } c \text{ not selected} \\ p_c(\delta_1^1, \delta_2^1), & \text{if channel c belongs to } \varphi_1 \\ p_c(\delta_1^2, \delta_2^2), & \text{if channel c belongs to } \varphi_2 \\ \max_k(p_c(\delta_1^k, \delta_2^k)), & \text{else} \end{cases},$$

where all the parameters are derived for two-source scenarios. Integration of the individual probabilities across all channels as done in Eq. (7) gives the conditional probability $p'(M, C \mid (\varphi_k)_k)$ for the current time frame assuming source φ_1 is the strongest. Then, the likelihood function is computed as follows:

$$p(M, C \mid (\varphi_k)_k) = K \alpha_n \max_P p'(M, C \mid P((\varphi_k)_k)), \quad (8)$$

where the maximization considers all the permutations P of the set $(\varphi_k)_{k=1,..,n}$ and α_n is used to adjust the relative strength between the zero-, one-, two- and three-source hypothesis. After training we fix α_n to the following values: $\alpha_2 = 0.0608$ and $\alpha_3 = 0.0907$.

Finally, we fix the probability of zero active source, i.e. $\boldsymbol{x} \in S_0$,

$$p(M, C \mid \boldsymbol{x}) = K \alpha_0, \quad \text{where } \alpha_0 = 10^{-24}. \quad (9)$$

3.4 HMM-based source tracking

The state space and the time axis are discretized and the standard Viterbi algorithm is employed to identify the optimal sequence of states. The algorithm attempts to reconstruct the initial tracks of the most probable sound sources in the scene where the maximum number of sources is fixed to three. Consequently, the decision of the system at every time frame includes the number of currently active sources and their estimated locations.

The computational expense of HMM algorithms can be reduced significantly by employing efficient implementations. The original tracks are continuous and thus pruning has been utilized to reduce the number of candidates to be examined for the current state. Also, beam search has been employed to reduce the state subspace considered in the evaluation of the current time frame. Finally, a corpus of 10 speech signals from the TIMIT database presented at a systematic set of spatial configurations is used for parameter estimation and training for the one-source and the two-source scenarios.

4. RESULTS

The performance of the tracking system presented in Section 3 is illustrated for a series of multiple moving source configurations. A pair of left and right signals is synthesized as described in Section 2 for a linear motion with constant speed.

Figure 1A shows the result of tracking one male and two female sources for a duration of 2.5 s. The sources move from left to right in a linear motion with respect to the virtual listener and with a constant speed of 1 m/s. The system is able to track the three sources across time, and indicates when a source is not active as long as it is not entirely masked by the interference.

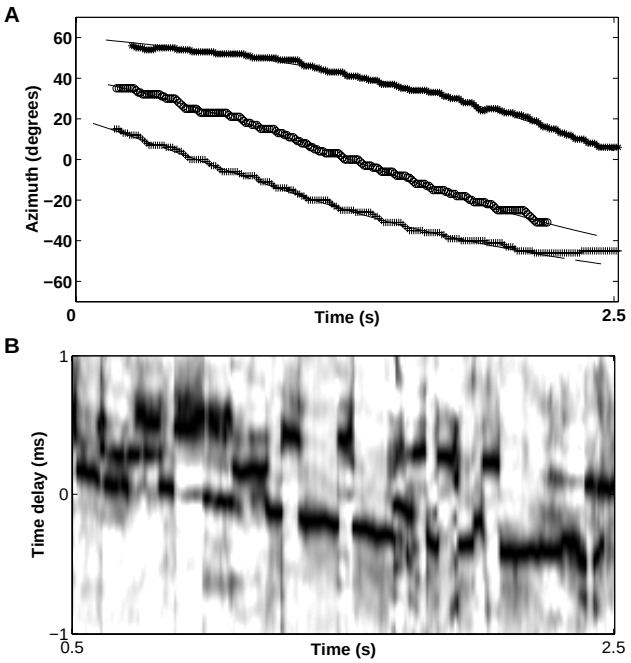

Figure 1. Tracking three moving sources. **A**. Continuous tracks obtained by applying the model. The solid lines give the original trajectories where a gap indicates a pause in the sentence. The '*', 'o' and '+' tracks correspond to the estimated tracks. **B**. Summarized cross-correlation across time.

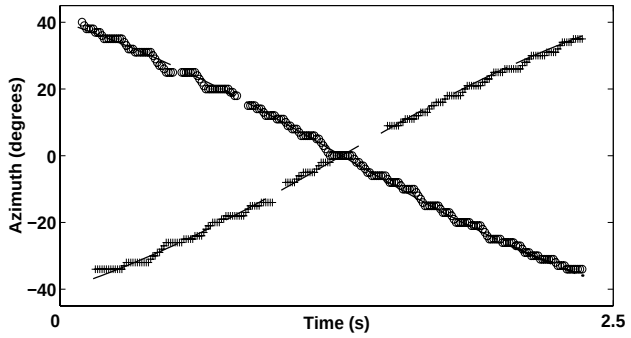

Figure 2. Source tracking for two intersecting sources. The solid lines give the original trajectories where a gap indicates a pause in the sentence. The '+' and 'o' tracks correspond to the estimated tracks.

For a comparison, we present in Fig. 1B the evolution of the standard cross-correlogram which is similar in principle to the generalized cross-correlation as used in [3]. Normalized cross-correlations for individual channels are summed and the results are shown for all time frames. Here the vertical axis shows the time lag in the plausible range from –1 ms to 1 ms and the darker regions correspond to stronger activities in the summarized cross-correlation. This representation utilizes only the ITD information and thus exhibits the multiple-peak problem. For an anechoic situation, the strongest peak is usually well correlated with the strongest source but the secondary peaks can be misleading. In fact, peaks associated with true sources are absent for considerable numbers of time frames. By combining ITD and IID in a statistical model and employing a tracking system, our model overcomes these problems.

A challenging task is shown in Fig. 2 where the two moving sources intersect each other. This scenario is obtained for a male speaker moving from left to right in front of the virtual listener with speed 1 m/s and a female speaker moving from right to left with the same speed. Our system is able to disambiguate the two tracks in this example. In general, the system needs to incorporate additional information to deal with intersecting motion tracks, e.g. spectral continuity and pitch continuity.

Similar results have been obtained for a variety of other configurations, including stationary sources.

5. CONCLUSION

We have proposed a new algorithm for tracking multiple moving sound sources. Our model integrates reliable information across time-frequency regions and imposes a continuity constraint through an HMM. Although the current system does not consider reverberation or complex motion trajectories, our framework is very promising for those situations also. Our study represents a first attempt to address auditory scene analysis with moving sound sources.

Acknowledgements. This research was supported in part by an NSF grant (IIS-0081058) and an AFOSR grant (F49620-01-0027). We thank M. Wu for many helpful discussions.

6. REFERENCES

[1] N. Roman, D. L. Wang and G. J. Brown, "Location-based sound segregation," *Proc. IEEE ICASSP*, 2002.

[2] D. E. Sturim, M. S. Brandstein and H. F. Silverman, "Tracking multiple talkers using microphone-array measurements," *Proc. IEEE ICASSP*, 1997.

[3] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," *Proc. IEEE ICASSP*, 2001.

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

[5] L. D. Stone, C. A. Barlow and T. L. Corwin, *Bayesian Multiple Target Tracking*, Artech House, 1999.

[6] M. Wu, D. L. Wang and G. J. Brown, "A multi-pitch tracking algorithm for noisy speech," to appear in *IEEE Trans. Speech Audio Processing*. An earlier version appears in *Proc. IEEE ICASSP*, 2002.

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

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

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

[10] 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. Am.*, vol. 91, pp. 1637-1647, 1992.

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

[12] 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.