SEPARATION AND TRACKING OF MULTIPLE SPEAKERS IN A REVERBERANT ENVIRONMENT USING A MULTIPLE MODEL PARTICLE FILTER GLIMPSING METHOD

Alireza Masnadi-Shirazi and Bhaskar D. Rao

Department of Electrical and Computer Engineering, University of California, San Diego
{amasnadi, brao}@ucsd.edu

ABSTRACT

In this paper, we explore the problem of separating and localizing multiple maneuvering speakers in a reverberant environment. Moreover, the speakers can become silent intermittently, making the problem more challenging especially when the speaker is moving while silent. A two stage method is proposed. In the first stage a multiple model particle filter (MMPF) is used to track the mixing matrices by "glimpsing" or listening in the silence gaps as the set of active speakers change with time. In the second stage, a secondary microphone array is utilized to localize and track the speakers by intersecting their directions of arrival (DOA) obtained from estimates of the tracked mixing matrices. Computer simulations using the proposed method are conducted showing improved performance when compared with an online independent component analysis (ICA)-based algorithm.

Index Terms— Multiple model particle filter, independent component analysis, source localization and tracking, blind source separation, direction of arrival

1. INTRODUCTION

The problem of separating original source signals from their multiple observed mixtures is commonly known as blind source separation (BSS). BSS of acoustical signals has a variety of applications including multimedia conferencing, high-quality hearing aids, automatic speech recognition of multiple concurrent speakers and background noise reduction for cell phones. Frequency-domain independent component analysis (ICA) is a prominent method of BSS for reverberant environments which works by transforming the mixtures to the frequency domain, performing regular ICA in each frequency bin and correcting for possible permutations in the frequency bins by using a post-processing method. Most ICA algorithms assume that the sources are spatially static, therefore resulting in a time-invariant mixing process. However, in most real world scenarios the sources can be moving around making the mixing time-varying. In order to deal with such a condition, an online or blockwise batch (quasi-online) form of BSS is necessary [1, 2].

Passive localization and tracking of multiple acoustical sources is also of great interest in the field of microphone arrays. Antonacci et al. first proposed a method that uses a quasi-online time-domain-based convolutive ICA method that attempts to (partially) separate the speakers and from that estimate the time difference of arrival (TDOA) for each speaker using the extrema of the demixing filters. The same is repeated for different microphone arrays in different positions in the room resulting in multiple TDOAs for each source. The location of the sources is then estimated by triangulation of the TDOA hyperbola locus points. However, because the sources could be permuted for each array and because they might have not been fully separated, TDOAs from the arrays can’t be matched to a specific source, resulting in false localizations. Therefore, more than one extra array (or auxiliary arrays using microphones within different array sets) is needed in order to reject such false positions. This localization method is accompanied by a particle filter on the source dynamics in order to track the sources [3, 4].

In this paper we intend to both fully separate and track the multiple speakers using an online algorithm. Sawada et. al has shown that for static sources, estimating correct mixing matrices in the frequency domain using ICA can lead to accurate estimations of the directions of arrival (DOA) [5]. The assertion of this paper is that for maneuvering sources, if one is able to track the mixing matrices accurately in the frequency domain to ensure full separation instead of partial separation, the accurate localization of the sources based on DOAs can be a straightforward consequence. This is in contrast to the aforementioned methods based on TDOA where the focus was on localization using already existing online BSS algorithms that might be able to only achieve partial separation. Moreover, we exploit a common form of temporal dynamics, especially present in speech, wherein the signals have silence periods intermittently, hence varying the set of active sources with time. By doing so we enable the algorithm to "glimpse" or listen in the gaps [6]. Utilizing such glimpsing strategy is essential in an online algorithm because if a source becomes silent but assumed active by the model, the update to the column of the mixing matrix corresponding to that source can

This research was supported by UC MICRO grants 07-034, 08-65 sponsored by Qualcomm Inc.
2. GENERATIVE MODEL

We assume there are \( L \) microphones in the array and \( M \) sources. After taking the short time Fourier transform (STFT) (with \( d \) frequency bins) of the convolutedly mixed (due to reverberation) signals corrupted with white Gaussian noise, the observations would end up having a linear mixture in each frequency bin \( k = 1, \ldots, d \) described as

\[
Y^{(k)}(t) = H^{(k)}(t)S^{(k)}(t) + W^{(k)}(t)
\]

where \( H^{(k)}(t) \) is the \( L \times M \) time varying mixing matrix for the \( k^{th} \) frequency bin. Since the noise is white it will have the same energy in all frequency bins. Hence the covariance of the noise can be written as \( \sigma_t = \text{diag}(\sigma_{w1}, \ldots, \sigma_{wt}) \). Throughout the rest of this paper all operations are carried out in each individual frequency bin, therefore the superscript \((k)\) is omitted for brevity. Each source is modeled as a multivariate GMM with \( C \) mixtures. The joint density of the sources is the product of the marginal densities, based on independency. Hence, we have

\[
P_S(S) = \prod_{j=1}^{M} \sum_{\ell=1}^{C} \alpha_{j\ell} G \left( S_j, 0, \sigma_{j\ell} \right)
\]

\[
= \sum_{q=1}^{GM} w_q G \left( S, 0, v_q \right)
\]

where \( \sum_{q=1}^{GM} = \sum_{\ell=1}^{C} \ldots \sum_{\ell=C}^{C} \), \( w_q = \prod_{j=1}^{M} \alpha_{j\ell} \) and \( v_q = \text{diag} \left( \sigma_{1\ell}, \ldots, \sigma_{M\ell} \right) \). The parameters \( \alpha_{j\ell} \) and \( \sigma_{j\ell} \) for \( j = 1, \ldots, M \) are fixed beforehand corresponding to a Gaussian mixture model (GMM) with zero means and varying variances (Gaussian scaled mixtures), hence having the shape of a symmetric multivariate super-Gaussian density [6]. Since each source can take on two states, either active or inactive, for \( M \) sources there will be a total of \( 2^M \) states. As a convention throughout this paper we will encode the states by a number between 1 and \( I = 2^M \) with a circle around it. Because we consider each bin separately, these states can vary across the frequency bins and indicate which column vector(s) of the mixing matrix is(are) present or absent for each bin.

Let the source indices form a set \( \Omega = \{1, \ldots, M\} \), then any subset of \( \Omega \) could correspond to a set of active source indices. For state \( x(t) = \Omega \), we denote the subset of active indices in ascending order by \( \Omega_i = \{\Omega_i(1), \ldots, \Omega_i(M_i)\} \subseteq \Omega \), where \( M_i \leq M \) is the cardinality of \( \Omega_i \). It can be easily shown that the observation density function for state \( \Omega \) is

\[
P_\Omega (Y(t)|H(t)) = \sum_{\Omega} w_{\Omega} G \left( Y(t), 0, A_{\Omega}(t) \right)
\]

where \( A_{\Omega}(t) = \sigma_t + H(t)^T v_{\Omega} H(t)^T \), \( \sum_{\Omega} = \sum_{\ell=1}^{C} \ldots \sum_{\ell=C}^{C} w_{\Omega} = \prod_{j=1}^{M_i} \alpha_{\Omega_i(j)\ell} \), \( H = [h_{\Omega_i(1)}, \ldots, h_{\Omega_i(M_i)}] \) being a subset of the full matrix containing only the \( \Omega_i^{(1)} \) \( \Omega_i\)th to \( \Omega_i(M_i) \) \( \Omega_i\)th columns and \( v_{\Omega} = \text{diag} \left( \sigma_{\Omega_i(1)\ell}, \ldots, \sigma_{\Omega_i(M_i)\ell} \right) \). When all the sources are active, the observation density in (3) uses the full mixing matrix and when none of the sources are active, the observation density reduces to white Gaussian noise [6].

We represent the evolution of the columns of the mixing matrices with indices \( m = 1, \ldots, M \) as a random walk model of

\[
h_{m}(t) = h_{m}(t-1) + \nu_{m}(t-1)
\]

where \( \nu_{m}(t) \) is a white Gaussian random variable with a diagonal covariance. We assume that the transition from one state to another follows a Markovian property with transition matrix

\[
\Pi = [\pi_{ij}] \text{ where } \pi_{ij} = Pr \left[ x(t) = i | x(t-1) = j \right].
\]

In the next section we describe a MMPF algorithm capable of tracking the mixing matrices and the sources’ activity pattern.

3. MULTIPLE MODEL PARTICLE FILTERING

Particle filtering is an online Bayesian state estimation technique widely used for nonlinear/nonGaussian state estimation. From Eqs. 1 and 3, it is clearly evident that the relationship between the observations and the states is nonlinear and nonGaussian. Particle filtering in junction with ICA for time-variant mixing has been proposed before for linear instantaneous mixing while assuming the sources were active at all times [7, 8]. In this section we describe a frequency domain MMPF for convolutive mixing capable of switching between states corresponding to different source activity patterns. Assuming \( N \) particles are used, the main steps of the MMPF is summarized as follows [9]:

1. Initialize the state particles \( \{h_{m}^n(0), m = 1, \ldots, M\}^N_{n=1} \) and \( \{x^n(0)\}^N_{n=1} \) based on an initial prior and using uniform weights \( \{w_{m}^n(0) = 1/N, m = 1, \ldots, M\}^N_{n=1} \) and \( \{x^n(0) = 1/N\}^N_{n=1} \).
2. Classify the particles to sets corresponding to different activity states, denoting \( n_i = \{ n | x^n_i(t) = \emptyset \} \) for \( i = 1, ..., I \). Next predict the new set of particles by drawing a new set of samples at time \( t \) according to state transitions described by
\[
\begin{align*}
\hat{h}_m^n(t) &= h_m^n(t-1) + v_m^n(t-1) \quad \text{state } m \text{ contains column } m \\
\hat{h}_m^n(t) &= h_m^n(t-1) \quad \text{state } m \text{ excludes column } m
\end{align*}
\]
for \( m = 1, ..., M \), \( i = 1, ..., I \) and \( n = 1, ..., N \). This model assumes that columns of the mixing matrices vary only when the corresponding sources are active. The reason for this is to avoid the particles from drifting when no information is available and the sources are silent. However, by keeping a memory of the silence patterns of the sources based on previous frames, the covariance of the cloud of particles can be increased virtually. This way the clouds of particles during the SBZs would be large enough to find the track once the sources become active again. After that if the sources remain active enough the covariances can be decreased back to normal (similar to recovering track in blind Doppler zones in radar signal processing [9]). Also, the activity pattern is predicted according to the rule that if \( x^n_i(t-1) = \emptyset \) then \( x^n_i(t) = \emptyset \) with probability \( \pi_{ji} \).

3. In this step the weights for the columns of the mixing matrices are updated as
\[
\begin{align*}
w_m^n_i(t) &= w_m^n_i(t-1)P(\emptyset | Y(t) | H^n_i(t)) \quad \text{state } m \text{ contains column } m \\
w_m^n_i(t) &= w_m^n_i(t-1) \quad \text{state } m \text{ excludes column } m
\end{align*}
\]
for \( m = 1, ..., M \), \( i = 1, ..., I \) and \( n = 1, ..., N \). The activity weights are updated as
\[
r^n_i(t) = r^n_i(t-1)P(\emptyset | Y(t) | H^n_i(t))
\]
for \( i = 1, ..., I \) and \( n = 1, ..., N \).

4. Normalize the activity weights so their sum is unit value
\[
r^n_i(t) \leftarrow \frac{r^n_i(t)}{\sum_n r^n_i(t)}
\]
and from that obtain the probability of each activity state
\[
p(x(t) = \emptyset | Y(1, ..., t)) = \sum_n r^n_i(t)
\]
The column weights are then normalized as
\[
w_m^n_i(t) \leftarrow w_m^n_i(t)p(x(t) = \emptyset | Y(1, ..., t))/\sum_n w_m^n_i(t)
\]
for \( m = 1, ..., M \), \( i = 1, ..., I \) and \( n = 1, ..., N \).

5. If the particles become degenerate resample them and reassign the weights to uniform.

6. Estimate the matrix columns using
\[
\hat{h}_m^n(t) = \sum_n w_m^n_i(t)h_m^n(t)
\]
and from that the sources can be reconstructed using a minimum mean square error (MMSE) estimator [6].

7. Permutation in the frequency bins is corrected using a correlation method on the activity patterns by keeping a memory of the past estimates of the sources in each frequency bin [10]. After that the sources are converted to the time domain.

4. LOCALIZATION AND TRACKING

Once an estimate of the time varying mixing matrices is found, the DOAs can be found using the method in [5]. If the same procedure is repeated in parallel for another microphone array placed at a different position in the room the sources can be located and tracked using triangulation (similar to a multiple bearings-only framework with static sensors in radar signal processing [9]). For simplicity we assume that the secondary DOA estimates are synchronous to the primary estimates and that zero delay transmission delay exists between the two. Because the DOA estimates can be jittering especially when the sources are silent for some time and suddenly become active, we propose to smooth the localization process of triangulation by incorporating kinematic dynamics for the motion of the sources. Therefore, another tracking stage is added where the DOA estimates are treated as measurements and the positions and velocities of the sources are treated as states. Because the relationship between the DOAs and the position of the sources is nonlinear [9], a method based on particle filtering is proposed again. Moreover, in order to model maneuvering sources a MMPF (similar to the tracking of the mixing matrices in the previous section) incorporating constant velocity and constant acceleration models in the \( x \) and \( y \) directions is employed. We note that because the algorithm is able to to first fully separate the sources and then localize them, the DOA estimates from the two arrays can be matched to a specific source by evaluating the correlation of the separated sources activity patterns [10], avoiding the problem of false positions due to possible permutations.

5. COMPUTER SIMULATIONS

The proposed algorithm was put to test in a simulated room settings using the image method. We assumed a room size of 8x5x3.5m with a reverberation time of 200ms. We picked the simple case of \( M = 2 \) speakers and \( L = 2 \) microphones for each array. The two arrays were placed facing each other.
The sampling frequency for speech sources was 8kHz and the spacing between the microphones in both arrays was 4cm in order to avoid spatial aliasing. The two sources moved in the same direction (one chasing the other) in a maneuvering radial pattern with an angular speed of around 6.4 deg/sec with respect to the primary microphone array. The total duration of the sources was 12.5 seconds with them being active only for an average of about 5.5 seconds. The data was corrupted with white Gaussian noise, with the noise level resulting in an input signal to noise ratio (SNR, in.) of 14(dB). Signal to disturbance ratio (SDR) is used as the performance measure for the separation phase. SDR is the total signal power of direct channels versus the signal power stemming from cross interference and noise combined. Number of particles used was N = 1000. Position root mean square error (RMSE) is used as a performance measure for the tracking phase.

In order to evaluate the results, the proposed method was compared to an online independent vector analysis (IVA) method with a normalized natural gradient nonholonomic constraint (NNGNC) [2]. Both algorithms were initialized the same way by performing batch IVA [6] on the first 2 seconds of data. The SDR using the proposed method measured to be 11.4 (dB) while the SDR of the online IVA algorithm came out to be 6.8 (dB). Fig. 1 shows the true positions of the sources along with the estimated positions using the proposed method. Fig. 2 compares the average position RMSE for the sources using the proposed algorithm and the online IVA algorithm. Because of the high jitter in the DOAs when using the online IVA method, tracking with a motion model failed to work. Therefore the positions found based on simple intersections of DOAs, without applying any motion model on the sources, was used to compute the RMSE of the online IVA algorithm. The audio files along with the video of the tracking phase are available at our website.\footnote{http://dsp.ucsd.edu/~ali/tracking/}

6. CONCLUSIONS

We have proposed a novel frequency domain particle filtering method capable of tracking the mixing matrices of maneuvering sources in a reverberant environment. A glimpsing approach is also incorporated to switch between different combinations of tracks when the source(s) become inactive, therefore avoiding losing track during such periods. The algorithm is also capable of recovering tracks during silence blind zones (SBZ) where the sources are moving while silent. Once the mixing matrices are correctly estimated, obtaining the directions of arrival (DOA) becomes a straightforward post-processing step. Using a secondary array positioned elsewhere in the room, the DOAs are matched up and triangulated by incorporating a multiple motion model on the source trajectories. Improved separation and tracking results were achieved in the simulations.

7. REFERENCES