MULTIMODAL BLIND SOURCE SEPARATION FOR MOVING SOURCES

S. M. Naqvi, Y. Zhang and J. A. Chambers

Advanced Signal Processing Group, Department of Electronic and Electrical Engineering
Loughborough University, Loughborough LE11 3TU, UK.
Email: {s.m.r.naqvi, y.zhang5, j.a.chambers}@lboro.ac.uk

ABSTRACT

A novel multimodal approach is proposed to solve the problem of blind source separation (BSS) of moving sources. The challenge of BSS for moving sources is that the mixing filters are time varying, thus the unmixing filters should also be time varying, which are difficult to track in real time. In the proposed approach, the visual modality is utilized to facilitate the separation for both stationary and moving sources. The movement of the sources is detected by a 3-D tracker based on particle filtering. The full BSS solution is formed by integrating a frequency domain blind source separation algorithm and beamforming: if the sources are identified as stationary, a frequency domain BSS algorithm is implemented with an initialization derived from the visual information. Once the sources are moving, a beamforming algorithm is used to perform real time speech enhancement and provide separation of the sources. Experimental results show that by utilizing the visual modality, the proposed algorithm can not only improve the performance of the BSS algorithm and mitigate the permutation problem for stationary sources, but also provide a good BSS performance for moving sources in a low reverberant environment.

Index Terms— BSS, multimodal signal processing, particle filtering, 3-D tracking, beamforming, FastICA.

1. INTRODUCTION

Most existing BSS algorithms are based on statistical information extracted from the received mixed data e.g. [1,2]. However, in many real applications, the sources may be moving, for example, a presenter may walk around inside a room. In such applications, there will be insufficient data length available, which limits the application of these algorithms. Thus BSS methods for moving sources are very important to solve the cocktail party problem in practice. Only a few papers have been presented in this area [3,4]. In [3] sources are separated by employing frequency domain ICA using a block-wise batch algorithm in the first stage, and the separated signals are refined by postprocessing in the second stage which constitutes crosstalk component estimation and spectral subtraction. In [4] they used an online PCA algorithm to calculate the whitening matrix and another online algorithm to calculate the rotation matrix, both algorithms are designed only for instantaneous source separation, and can not separate convolutive mixed signals. Fundamentally, it is very difficult to separate convolutive mixed signals by utilizing the statistical information only extracted from audio signals.

Work supported by the Engineering and Physical Sciences Research Council (EPSRC) of the UK (EP/C535308/2).

In this work, a multimodal approach is proposed by utilizing not only received linearly mixed signals, but also the video information obtained from video cameras and a key component in the proposed approach is the tracking of speakers. A video system can capture the approximate positions and velocities of the speakers, from which we can identify the directions and motions, i.e., stationary or moving speakers. If the source is stationary a geometrically based initialization is performed to improve the performance of the frequency domain BSS algorithm and mitigate the permutation problem. In the case of moving sources a beamforming method is used to enhance the signal from one source direction and reduce the energy received from another source direction, so that source separation can be obtained. Although the beamforming approach can only reduce the signal from a certain direction and the reverberance of the interference still exists, it can obtain a good separation performance in a low reverberation environment (reverberation time (RT) is 130ms). Performing BSS in rooms with large RT$>$130ms remains as a research challenge. Note that the beamforming approach only depends on the direction of source signals, thus an online real time source separation can be obtained.

The paper is organized as follows: Section-II presents the system model, Section-III describes the source separation by combining frequency domain BSS and beamforming and experimental results are provided in Section-IV based on real room recordings. Finally, in Section-VI we conclude the paper.

2. THE SYSTEM MODEL

The schematic diagram of the system is shown in Figure 1. The proposed approach can be divided into two stages: human tracking to obtain position and velocity information and source separation by utilizing the position and velocity information based on frequency domain BSS algorithms and beamforming.

The static video cameras are calibrated off line which recovers calibration parameters i.e. the interior orientation, the exterior orientation and translation vector, the power series coefficients for distortion, and image scale factor. Video cameras are synchronized by the external hardware trigger module and frames are captured at the rate of $f_v = 25$ frames/sec, which means $T_v = 1/25$sec. We extract the face of each speaker in the images of synchronized video cameras to find the position of each speaker at each state (time). In each image frame this is performed on the basis of a skin model and matching the image with a face model. The position of the lips of a speaker is determined from the extracted face region in image coordinates $\boldsymbol{\iota}_c = [x, y]^T$. With the help of the above calibration parameters we calculate the image coordinates of each speaker in 3-D world coordinates to get the real world position of each speaker $\mathbf{z}$ in $\mathbb{R}^3$. The calculated position of the lips $\mathbf{z}$ of each speaker is then

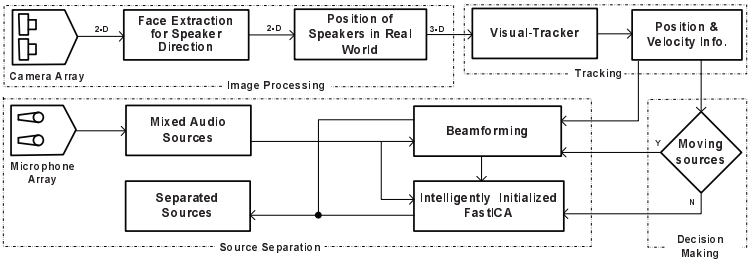

Fig. 1. System block diagram: Face extraction is based on a skin model and template matching with the human face, this 2-D image information of the video cameras is converted to 3-D world co-ordinates through the calibration parameters. The 3-D estimates are fed to the visual-tracker, and on the basis of position and velocity information from the tracking, the sources are separated either by beamforming or by intelligently initializing the FastICA algorithm.

used a measurement in a particle filter based tracker, and the position and velocity obtained from the tracker will be used in the source separation.

2.1. Tracking the Source Position and Velocity

The 3-D visual tacker is based on particle filters (for detail see [5]) and here we will only provide the state and measurement model.

The state and measurement configurations are $\mathbf{x}_{0:k}, \mathbf{z}_{0:k} = \{\mathbf{x}_j, \mathbf{z}_j, j = 0, ..., k\}$, where $\mathbf{x}_{0:k}$ formulates the state sequence of the target which we want to obtain, and $\mathbf{z}_{0:k}$ is the observation sequence, both in $\mathbb{R}^3$. For each iteration, the target state evolves according to the following discrete-time stochastic model:

$$\mathbf{x}_k = \mathbf{f}_k(\mathbf{x}_{k-1,k-2}, k) + \mathbf{v}_{k-1} \tag{1}$$

where $\mathbf{f}_k(\mathbf{x}_{k-1,k-2}, k) = 2\mathbf{x}_{k-1} - \mathbf{x}_{k-2}$ represents a random walk model for the state $\mathbf{x}_k$ and is used for the approach and k is the discrete index. Process noise $\mathbf{v}_{k-1}$ is white noise, generally non-Gaussian and caters for under modelling effects and unforseen disturbances in the state model, and its covariance matrix is Q_v.

The objective of the filter is to estimate recursively state $\mathbf{x}_k$ from the measurement $\mathbf{z}_k$ and the measurement equation is:

$$\mathbf{z}_k = \mathbf{h}_k(\mathbf{x}_k, k) + \mathbf{r}_k \tag{2}$$

where $\mathbf{h}_k(\mathbf{x}_k, k) = \mathbf{x}_k$ and $\mathbf{r}_k$ is a vector of Gaussian random variables with covariance matrix Q_r which caters for the measurement errors.

The output of the 3-D visual tracker is position $\mathbf{x}_k = [x_k^x, x_k^y, x_k^z]^T$ and velocity s_k of a speaker at each state k. The distance between consecutive states is calculated as $d_k = ||\mathbf{x}_k - \mathbf{x}_{k-1}||_2$ and the velocity at state k is calculated as $s_k = d_k/T_v$ where $||.||$ denotes the Euclidean norm. The change in the position of a speaker with respect to the previous state plays a critical role in source separation either by using beamforming or by intelligently initializing the FastICA.

2.2. Source Separation

The audio mixtures from the microphone sensor array are separated with the help of visual information from the 3-D tracker. On the basis of velocity information if the sources are stationary for a certain period ($T_k = 5$sec in our simulations) we separate the sources with intelligently initialized FastICA, otherwise, we separate the sources by beamforming. The other important parameter to be calculated before starting the source separation is the angle of arrival of each speaker to the sensor array. By having the position information of the microphones and the speakers at each state from the 3-D visual tracker we can easily calculate the angle of arrival $\boldsymbol{\theta}_{0:k}$ of speakers to the microphone sensor array.

In the intelligent office where our recordings are taken the microphones used are uni-directional. By using a short-time discrete Fourier transform (DFT) the mixing process can be formulated as follows: having M statistically independent real sources a multi-channel FIR filter $\mathbf{H}(\omega)$ producing N observed mixed signals can be described as (we assume there is no noise or noise can be deemed as a source signal in the model for simplicity)

$$\mathbf{u}(\omega) = \mathbf{H}(\omega)\mathbf{s}(\omega) \tag{3}$$

where $\mathbf{H}(\omega) = [\mathbf{h}_1(\omega), ..., \mathbf{h}_M(\omega)]$ and the source separation can be described as

$$\mathbf{y}(\omega) = \mathbf{W}(\omega)\mathbf{u}(\omega) \tag{4}$$

where $\mathbf{W}(\omega) = [\mathbf{w}_1(\omega), ..., \mathbf{w}_N(\omega)]$. In this work to demonstrate the proposed approach we consider the exactly determined convolutive BSS problem i.e. $N = M = 2$, without loss of generality. The unmixing matrix $\mathbf{W}(\omega)$ for each frequency bin is formulated as

$$\mathbf{W}(\omega) = inv(\mathbf{H}(\omega)^H) \tag{5}$$

where $inv(.)$ is inverse of the matrix and $(.)^H$ is Hermitian transpose. The delay vector $\mathbf{h}_1(\omega)$ is formulated as

$$\mathbf{h}_1(\omega) = [1, \ldots, e^{j(N-1)d\cos(\theta)\omega/c}]^H \tag{6}$$

where d is the distance between the sensors and c is the speed of sound in air. Ideally, $\mathbf{h}_1(\omega)$ should be the sum of all echo paths, which are not possible to be tracked, therefore it is approximated by neglecting the room reverberations.

Finally, by placing $\mathbf{W}(\omega)$ in (4) we estimate the sources. Since the scaling is not a major issue [6] and there is no permutation problem, therefore we can align the estimated sources for reconstruction in the time domain.

If the sources are stationary for time T_k we initialized the FastICA [7] with the above $\mathbf{H}(\omega)$ similarly to as in [2].

3. EXPERIMENTS AND RESULTS

Data are collected in a 4.6 x 3.5 x 2.5 m^3 intelligent office. Video cameras are fully synchronized with external hardware trigger module and frames are captured at $f_v = 25$Hz with an image size of 640x480 pixels. Both video cameras have overlapping field of view. The duration between consecutive states is $T_v = 1/25sec$. Audio

recordings are taken at $f_a = 8$KHz and are synchronized manually with video recordings. Distance between the audio-sensors is $d = 4cm$. Skin models for the people in recordings were developed off line. The other important variables are selected as: number of sensors and speakers $N = M = 2$, the number of particles was $N_p = 600$ and results were obtained using 4 runs, the number of images is $k = 525$ which indicates $21sec$ of data, $T_k = 5sec$, $Q_v = 10^{-4}I$, $Q_r = 10^{-2}I$, FFT length $T = 2048$ and filter length $Q = 1024$, and the room impulse duration is $RT = 130$ms. In our proposed algorithm we use non-linearity for FastICA $G(y) = \log(b + y)$, with $b = 0.1$. In the experiments speaker 1 is stationary and speaker 2 is moving around a table in a tele-conference scenario.

3.1. 3-D Tracking and Angle of Arrival Results

In this section we will discuss the results obtained from tracking. Since speaker 2 is moving around the table (speaker 1 is stationary) so we will discuss the tracking results of the speaker 2 in detail. Since we have colour video cameras therefore the face detection of the speakers is possible by using the skin model as discussed in Section-2. In Figure 2 the colour blob indicates that the faces are detected well. Since in the dense environment as shown in Figure 2 it is very hard to detect the lips directly, we approximate the center of the detected face region as the position of the lips in each sequence.

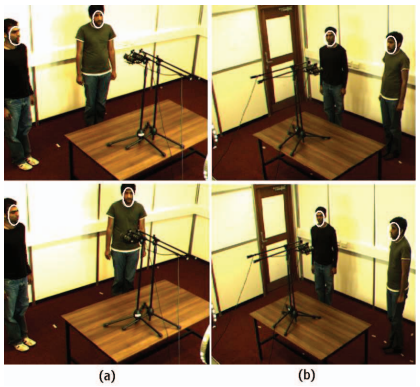

Fig. 2. 3-D Tracking results: frames of synchronized recordings, (a) frames of first camera and (b) frames of second camera; face detection based on skin model and template matching efficiently detected the faces in the frames.

The approximate 2-D position of the lips of the speaker in both synchronized camera frames at each state is converted to 3-D world coordinates. With this measurement we update the particle filter and results of the 3-D tracker are shown in Figure 3. The gait of the speaker is not smooth and the speaker is also stationary for a while at some points during walking around the table which provides a good test for the evaluation of 3-D tracker as well as for source septation methods, and it is also clear in the 3-D tracking results shown in Figure 3.

In order to view the tracking results in more detail, we plotted the tracking results in xy axes. Figure 4 clearly shows that the error in detection and conversion (measurement error) is almost corrected by the particle filter. Since the speakers and microphones are approximately at the same level therefore in the results of tracking Figure 5 we find that the effective movement of the speaker 2 was in the x and y-axis therefore the effective change in the angle of arrival was only in the xy plane. The angle of arrival of speaker 1 is 51.3 $degrees$ and the angles of arrivals of speaker 2 are shown in Figure 6.

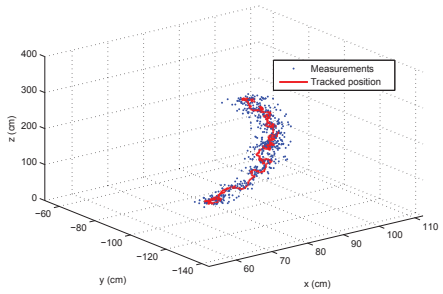

Fig. 3. 3-D Tracking results: PF based 3-D tracking of the speaker while walking around the table in the intelligent office.

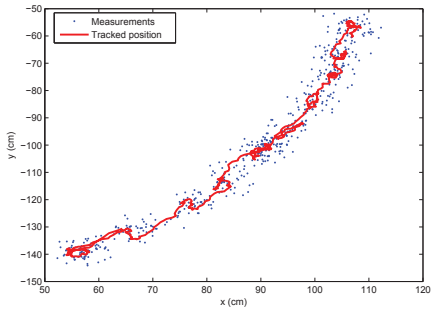

Fig. 4. 3-D Tracking results: PF based tracking of the speaker in the x and y-axis, while walking around the table in the intelligent office. The result provides more in depth view in the x and y-axis.

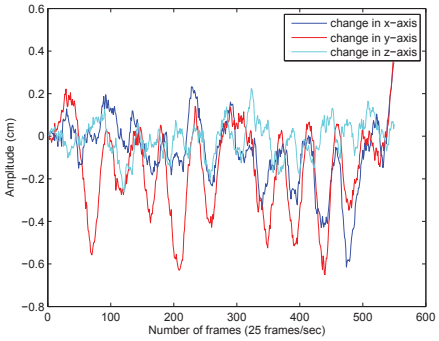

Fig. 5. 3-D Tracking results: PF based tracking of the speaker. The result provides the information which helps in deciding the method to separate the sources either by beamforming or by FastICA.

3.2. BSS Results

If the sources are moving we separate the sources by the beamformer and when the sources are stationary we separate the sources by intelligently initializing FastICA (IIFastICA). For the stationary case recorded mixtures of length of 5sec are separated (*for objective evaluation we convolved the signals with recorded real room impulse response, and separation of real room recordings are evaluated subjectively by listening tests*) and results are shown in Figure 7. The data length of the mixtures used for the beamforming case is 0.4sec

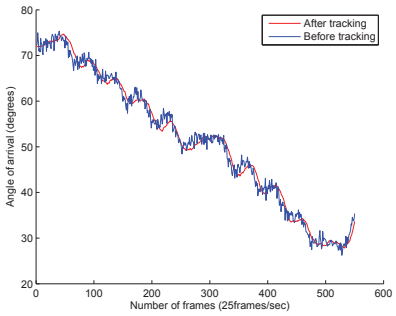

Fig. 6. Angle of arrival results: Angle of arrival of the speaker 2 to the sensor array. The estimated angle before tracking and corrected angle by PF are shown. The change in angle is not smooth because of the gait of the speaker.

(near to the moving case and for comparison otherwise beamforming is independent of data length) and the results are shown in Figure 8. The resulting performance indices [2] in the top of Figures 7,8 show good performance i.e. close to zero across the majority of the frequency bins. We also evaluate permutation on the basis of the criterion mentioned in [2] i.e. $[abs(G_{11}G_{22}) - abs(G_{12}G_{21})] > 0$ mean no permutation, where system matrix $G = WH$ and G_{ij} is the ijth element. In the bottom of Figures 7,8 the results confirm that the proposed algorithm automatically mitigates the permutation at each frequency bin. Since there is no permutation problem therefore sources are finally aligned in the time domain. The improvement in signal-to-interference ratio (SIR-Improvement) [6] for IIFastICA is $12.5dB$ and for beamforming is $9.5dB$.

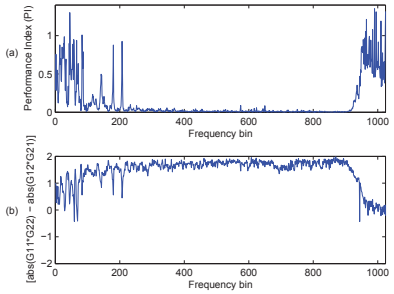

Fig. 7. BSS Results: performance index at each frequency bin for proposed IIFastICA algorithm at the top and evaluation of permutation at the bottom. A lower PI refers to a superior method and $[abs(G_{11}G_{22}) - abs(G_{12}G_{21})] > 0$ means no permutation.

Finally, separation of real room recordings were evaluated subjectively by listening tests, six people participated in the listening tests and mean opinion score is provided in Table 1 (MOS tests for voice are specified by ITU-T recommendation P.800).

4. CONCLUSIONS

In this paper a new multimodal BSS approach is proposed to solve the moving source separation problem. Video information is utilized which provides velocity and direction information of the sources.

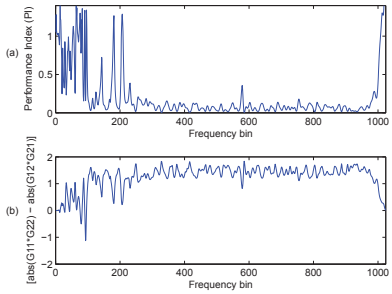

Fig. 8. BSS Results: performance index at each frequency bin for 3-D tracking based angle of arrival information used in beamforming at the top and evaluation of permutation at the bottom. A lower PI refers to a superior method and $[abs(G_{11}G_{22}) - abs(G_{12}G_{21})] > 0$ means no permutation.

Table 1. Subjective evaluation: MOS for separation of real room recordings, by intelligently initialized FastICA (IIFastICA) when sources are stationary, and by beamforming when sources are moving.

Algorithms	IIFastICA	Beamforming
Mean opinion score	4.7	3.8

The direction information is then utilized to facilitate the beamforming and source separation. As shown by the simulation results, the proposed approach has a good performance for both stationary and moving sources, which is not previously possible. This work provides a substantial step forward towards the solution of the real cocktail party problem.

5. REFERENCES

[1] W. Wang, S. Sanei, and J.A. Chambers, "Penalty function based joint diagonalization approach for convolutive blind separation of nonstationary sources," *IEEE Trans. Signal Processing*, vol. 53, no. 5, pp. 1654–1669, 2005.

[2] S. M. Naqvi, Y. Zhang, T. Tsalaile, S. Sanei, and J. A. Chambers, "A multimodal approach for frequency domain independent component analysis with geometrically-based initialization," *Proc. EUSIPCO, Lausanne, Switzerland*, 2008.

[3] R. Mukai, H. Sawada, S.Araki, and S. Makino, "Robust real-time blind source separation for moving speakers in a room," *Proc. IEEE ICASSP, Hong Kong*, 2003.

[4] K. E. Hild-II, D. Erdogmus, and J. C. Principe, "Blind source extraction of time-varying, instantaneous mixtures using an online algorithm," *Proc. IEEE ICASSP, Orlando, Florida, USA*, 2002.

[5] B. Ristic, S. Arulampalam, and N. Gordon, *Beyond the Kalman Filter: Particle Filter for Tracking Applications*, Boston—London: Artech House Publishers, 2004.

[6] S. Sanei, S. M. Naqvi, J. A. Chambers, and Y. Hicks, "A geometrically constrained multimodal approach for convolutive blind source separation," *Proc. IEEE ICASSP*, pp. 969–972, 2007.

[7] E. Bingham and A. Hyvärinen, "A fast fixed point algorithm for independent component analysis of complex valued signals," *Int. J. Neural Networks*, vol. 10, no. 1, pp. 1–8, 2000.