A Multimodal Approach to Blind Source Separation of Moving Sources

Syed Mohsen Naqvi, Miao Yu, and Jonathon A. Chambers, *Senior Member, IEEE*

***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 calculate 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 video cameras. Positions and velocities of the sources are obtained from the 3-D tracker based on a Markov Chain Monte Carlo particle filter (MCMC-PF), which results in high sampling efficiency. The full BSS solution is formed by integrating a frequency domain blind source separation algorithm and beamforming: if the sources are identified as stationary for a certain minimum period, a frequency domain BSS algorithm is implemented with an initialization derived from the positions of the source signals. Once the sources are moving, a beamforming algorithm which requires no prior statistical knowledge is used to perform real time speech enhancement and provide separation of the sources. Experimental results confirm that by utilizing the visual modality, the proposed algorithm not only improves the performance of the BSS algorithm and mitigates the permutation problem for stationary sources, but also provides a good BSS performance for moving sources in a low reverberant environment.**

***Index Terms*—Beamforming, blind source separation (BSS), FastICA, Markov Chain Monte Carlo (MCMC) particle filtering, multimodal signal processing, 3-D tracking.**

I. Introduction

BLIND source separation (BSS) of acoustic signals is a challenging problem when applied to a real environment, such as within an office occupied by a number of speakers, and remains a topic of considerable active research due to many potential applications [1]. BSS consists of estimating sources from such observed audio mixtures with only limited information and the associated algorithms have been conventionally developed in either the time or frequency domains [2]–[16]. Frequency-domain convolutive blind source separation (FDCBSS) has however been a more popular approach as the time-domain convolutive mixing is converted into a number of independent complex instantaneous mixing operations. The permutation problem inherent to FDCBSS presents itself when reconstructing the sources from the separated outputs of these instantaneous mixtures. It is more severe and destructive than for time-domain schemes as the number of permutations grows exponentially with the number of instantaneous mixtures [10].

Most existing BSS algorithms assume that the sources are physically stationary, i.e., the mixing filters are fixed. All these algorithms are based on statistical information extracted from the received mixed data [3]–[5]. 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 generally be insufficient data length available over which the sources are physically stationary, which limits the application of these algorithms. Thus BSS methods for moving sources are very important to solve the cocktail party problem in practice [17]. Only a few papers have been presented in this area [18]–[24]. In [18], 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 the case of [19], they used a framewise online algorithm in the time domain. However, both these two algorithms potentially assume that in a short period the sources are physically stationary, or the change of the mixing filters is very slow, which are very strong constraints. In [21], BSS for time-variant mixing systems is performed by piecewise linear approximations. In [22], they used an online PCA algorithm to calculate the whitening matrix and another online algorithm to calculate the rotation matrix. However, both algorithms are designed only for instantaneous source separation, and cannot separate convolutive mixed signals. In [24], it is assumed that mixing process is changing sufficiently slowly, so that one can find a window length that is short enough that the mixing can reasonably be approximated as stationary. Fundamentally, it is very difficult to separate convolutively mixed signals by utilizing the statistical information only extracted from audio signals, and this is not the manner in which humans solve the problem [25] since they generally use both their ears and eyes.

In this paper, a multimodal approach is therefore proposed by utilizing not only received linearly mixed signals, but also video information obtained from cameras. 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, of the speakers. A source is identified as stationary if its velocity is approximately zero for a certain minimum period, so that enough data length can be obtained for frequency domain BSS algorithms. Furthermore, the direction of the source signals

Manuscript received October 12, 2009; accepted December 11, 2009. Date of publication July 8, 2010; date of current version September 15, 2010. This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) of the U.K. under Projects EP/C535308/2 and EP/H049665/1. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Maurizio Omologo.

The authors are with the Advanced Signal Processing Group, Department of Electronic and Electrical Engineering, Loughborough University, Loughborough, Leicestershire LE11 3TU, U.K. (e-mail: s.m.r.naqvi@lboro.ac.uk; m.yu@lboro.ac.uk; j.a.chambers@lboro.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2010.2057198

can also be obtained from the video cameras, and a geometrically based initialization can then be performed to improve the performance of the frequency domain BSS algorithm and mitigate the permutation problem [8]. If the velocity is larger than an upper bound value, the source is identified as moving. In this case, a beamforming method which does not need prior statistical information, in common with the fundamental assumptions in blind source separation, 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, which are also limitations in the BSS approach, it can obtain an acceptable separation performance in a low reverberant environment. Note that the beamforming approach only depends on the direction of the source signals, and no received audio data are required, thus an online real time source separation can be obtained [26].

The remainder of this paper is organized as follows. Section II presents the related work, Section III provides the system model, and Section IV explains the tracking process. Section V describes the source separation by combining frequency domain BSS and beamforming. Experimental results are provided in Section VI based on real room recordings from our intelligent office. Finally, in Section VII we conclude the paper.

II. Related Work

Most existing BSS algorithms are based on the statistical information, second-order statistics (SOS)/higher order statistics (HOS), extracted from the recorded data. Such methods are generally not applicable in CBSS of moving sources due to data length limitations and are therefore not included for comparisons purposes in our simulation studies with moving sources. In the context of CBSS of moving sources in a moderate reverberant environment, with a reverberation time (RT) $<$130 ms, we believe that a multimodal approach is necessary which exploits different processing techniques as a function of the velocity of the speakers. A key component in this approach is the tracking of speakers. Many methods have been proposed for the tracking of speakers on the basis of audio information, visual information, or audio-visual fusion [27]–[45]. Broadly speaking, the differences among the existing approaches arise on the basis of single-person or multi-person tracking and the type of sensor configuration used. In most of the works [27], [28], [32], [33], [35], [42], on the basis of simple sensor configuration, either a single person is tracked in a single-person scene or the current active speaker is tracked in the multi-person scene. Multi-person tracking has been studied in [36], [38], [39], [41], [43] on the basis of only a single modality, either audio or video. In more recent works [29], [30], [37], [40] the multi-person tracking problem has been studied by using the audio-visual sensor configuration. To the best of our knowledge, the most recent work on tracking, near to our requirement, is proposed by Gatica-Perez [40]. In this paper, a detect before track technique is applied, and a small microphone array with multiple uncalibrated cameras with non-overlapping field of view (FOV) is used for sensor configuration. For detection, audio observations are derived from a source localization algorithm and visual observations are based on models of the shape and spatial structure of human heads. For tracking, a 2-D tracker in the image plane is implemented with a Markov Chain Monte Carlo particle filter (MCMC-PF). In our case, for source separation, 3-D positions of the speakers are required to handle complicated human motions. Therefore, initially, video cameras should be calibrated [46] and have overlapping FOVs, because at least two cameras are required for conversion of 2-D image coordinates to 3-D real-world coordinates. Second, it is computationally better to use one 3-D tracker rather than two 2-D trackers. Finally, audio localization is not effective due to the complexity in the case of multiple concurrent speakers. Localization for a single active speaker based only on audio is also difficult because human speech is an intermittent signal and contains much of its energy in the low-frequency bins where spatial discrimination is imprecise, and locations estimated only by audio are also affected by noise and room reverberations [47]. In [47], the tracker proposed in [40] is implemented for speech enhancement and the simulation results confirm that for stationary speakers and overlapping speech utterances the audio-visual localization improves by 2 cm and 3 cm, respectively, as compared to using only visual information. McCowan in [48] proposed that any time when the distance between the tracked speaker location and the focus location of the beamformer exceeds 5 cm, the beamformer channel filters should be recalculated, so practically there is no significant improvement by integrating audio localization. In other recent works [39], [44] only audio information is used. In [44], particle filtering is used for acoustic source localization and it is assumed that a single acoustic source with known speed of wave propagation is present in a reverberant environment. In [39], time difference of arrival (TDOA) estimation and localization of moving speakers are proposed (near to our requirement) which distinguish individual speakers in a multipath environment by associating one TDOA per frame to the predominant speaker. In the situation when speakers are simultaneously speaking and moving, both the above methods have limitations. In [49], joint acoustic source localization and orientation estimation using sequential Monte Carlo is presented and it is also highlighted in the paper that in a situation where only one microphone pair (sensor configuration used in this work) provides measurements then the performance is predictably poorer. In another recent work [45], audio and visual information is used for tracking of a speaker in a cluttered indoor environment and localization based on audio is discussed for only a single active speaker at a time. Therefore, in the proposed approach we track the speakers by using only visual information motivated by Colin Cherry's observation that the human approach to solve the cocktail party problem exploits visual cues [17], [50]. In our application environment, an intelligent office, the cameras also benefit from being mounted above the height of a human and thereby make it easier to discriminate sources in close spatial proximity. In the proposed approach, the source localization is performed by using the state-of-the-art Viola-Jones face detector [51]. The 3-D visual tracker is implemented with an MCMC-PF which results in high sampling efficiency. We stress that the domain of the proposed approach in this paper lies in system integration and the main contribution is to provide the

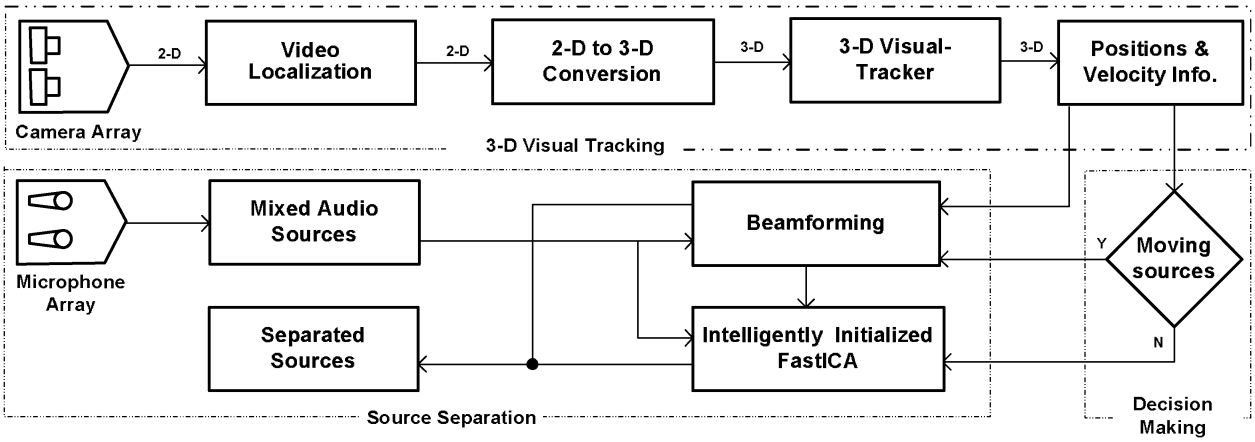

Fig. 1. System block diagram. Video localization is based on state-of-the-art Viola-Jones face detector [51], two fully calibrated color video cameras are used to determine the approximate 2-D positions of the speakers. The 2-D image information of the two video cameras is converted to 3-D world coordinates through the calibration parameters and optimization method. The approximated 3-D locations are fed to the visual-tracker, and on the basis of estimated 3-D real world position and velocity from the tracking, the sources are separated either by beamforming or by intelligently initializing the FastICA algorithm.

proof of the concept for CBSS of moving sources. The areas of detection and tracking are disciplines in their own rights and we simply exploit recent results from these fields to provide geometric information to facilitate a novel multimodal approach to CBSS. The output of the tracking is position and velocity information, on the basis of which we divide source separation into two parts to provide the full BSS solution. As will be shown in later simulations, the proposed approach can provide a reasonable BSS performance for moving sources in a low reverberant environment in which the RT is 130 ms. Performing BSS in rooms with large RT typically $>$130 ms remains as a research challenge.

III. System Model

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 or beamforming. The schematic diagram of the system is shown in Fig. 1.

For the localization of the sources we use two fully calibrated color video cameras to determine the approximate positions of the speakers. Both cameras are calibrated by the Tsai calibration (non-coplanar) technique [46] and synchronized by the external hardware trigger module and frames are captured at the rate of $f_v = 25$ frames/s, which means $T_v = 1/25$ s. We extract the face of each speaker in the images of both cameras to find the position of each speaker i at each state (time) k. In each image frame, the face can be extracted by the state-of-the-art Viola-Jones face detector [51]. It is highlighted that for this proof of concept work we assume the full face of a speaker is clearly visible and a simple geometric visual cue, i.e., the center of the face is available. The machine cocktail party problem is very challenging and our work is only to approach the ability of a human to solve this task. It is easy to contrive situations where a human would fail in this task and these are beyond the scope of this work. Further details are given in Section IV-B.

It is common in many science and engineering situations to estimate the hidden state of a system that changes over time using a sequence of noisy observations made on the system. Normally, the state-space approach, which focuses attention on the state vector of the system, is adopted for modeling a dynamic system. In this approach the 3-D location of each speaker is estimated by using the Bayesian multispeaker state space approach. The 3-D multispeaker observation is defined as $\mathbf{Z}_{1:k} = \{\mathbf{Z}_{1,1:k}, \ldots, \mathbf{Z}_{n,1:k}\}$, where $\mathbf{Z}_{i,1:k}$ represents the observations of speaker i and the multispeaker state configuration is defined as $\mathbf{X}_{1:k} = \{\mathbf{X}_{1,1:k}, \ldots, \mathbf{X}_{n,1:k}\}$. The filtering distribution of states given observations $p(\mathbf{X}_k|\mathbf{Z}_{1:k})$ is recursively approximated using an MCMC particle filter and the algorithm is explained in Section IV.

After estimating the 3-D position of each speaker the velocity information is extracted, if the sources are physically stationary for a certain period T_k, then the positions of the speakers are incorporated within the Intelligently Initialized FastICA (IIFastICA) algorithm otherwise they are used within the beamformer to obtain the source separation for stationary or moving sources. The details of the beamformer and IIFastICA are explained in Section V.

IV. 3-D Visual Tracker

The most suitable candidate for a 3-D multispeaker visual tracker is a particle filter because the probabilistic state-space formulation (non-Gaussian) and the requirement for the update of information on receipt of new measurements are ideally suitable for the Bayesian approach, which provides a rigorous general framework for dynamic state estimation problems. In the Bayesian approach to stochastic state estimation, the idea is to construct the posterior probability density function (pdf) of the state based on all the available information, including the received observations. Since such a pdf contains all the available statistical information, it can be considered to be the complete solution to the estimation problem.

For many problems, some sort of recursive processing is required in that at each time an observation is received, an estimate is required based on that observation. This may be achieved by the use of a recursive filter. Essentially, such a filter comprises of prediction and update stages. During the prediction stage, the

state pdf is predicted using the state model. Since the state is usually subject to some unknown disturbances (modeled as random noise), prediction generally deforms the state pdf. The predicted pdf, resulting from the prediction stage, is modified by the latest observation during the update stage. The update operation is achieved through Bayes' rule. The advantage of this recursive filtering is that the received data can be processed sequentially rather than as a batch. The posterior density $p(\mathbf{X}_k|\mathbf{Z}_{1:k})$ is recursively calculated by Bayes' rule according to

$$p(\mathbf{X}_k|\mathbf{Z}_{1:k}) \propto p(\mathbf{Z}_k|\mathbf{X}_k) \int p(\mathbf{X}_k|\mathbf{X}_{k-1}) p(\mathbf{X}_{k-1}|\mathbf{Z}_{1:k-1}) d\mathbf{X}_{k-1} \tag{1}$$

where $p(\mathbf{X}_k|\mathbf{X}_{k-1})$ denotes the multispeaker state model and $p(\mathbf{Z}_k|\mathbf{X}_k)$ represents the multispeaker measurement model. In general, no closed-form solution exists for (1) although these recursions can be approximated by Monte Carlo simulations of a set of particles having associated discrete probability mass and the generic particle filter is described in [52]. A particle filter recursively approximates the filtering distribution $p(\mathbf{X}_k|\mathbf{Z}_{1:k})$ by a weighted set of N_p particles at time k, $\{\mathbf{X}_k^n, \omega_k^n\}_{n=1}^{N_p}$, by using the weighted particles at the previous time-step $k-1$, $\{\mathbf{X}_{k-1}^n, \omega_{k-1}^n\}_{n=1}^{N_p}$, and the new update will be

$$p(\mathbf{X}_k|\mathbf{Z}_{1:k}) \approx K^{-1} p(\mathbf{Z}_k|\mathbf{X}_k) \sum_{n=1}^{N_p} \omega_{k-1}^n p\left(\mathbf{X}_k|\mathbf{X}_{k-1}^n\right) \tag{2}$$

where K is a normalization constant, $(.)^n$ refers to the n_{th} particle, and N_p is the number of particles, so that we have a discrete approximation of the true posterior. As N_p approaches to infinity, this discrete formulation will converge to the true posterior distribution. However, practically it is impossible to sample infinite number of samples from any distribution and the posterior distribution $p(\mathbf{X}_k|\mathbf{Z}_{1:k})$ to be estimated is also not in a closed form. The sampling mechanism is discussed in the sequel.

The three important items of the probabilistic multispeaker 3-D visual tracker, the state model, the measurement model and the MCMC-sampling mechanism are formulated in the following three subsections.

A. State Model

There are several state models that can be used to represent the state transition. In [53], we used the random walk model; another model which is shown to work well, to represent the time-varying location of a speaker in a typical room [53], [54], is the Langevin model [55], also used in [31], [42], and [44]. The motion of the speakers in each coordinate is assumed to be independent in this state model. In the $x-$coordinate this motion is described as

$$\begin{aligned} \dot{x}_k &= a_x \dot{x}_{k-1} + b_x F_x \\ x_k &= x_{k-1} + \Delta T \dot{x}_k \\ a_x &= e^{-\beta_x \Delta T} \\ b_x &= v_x \sqrt{1 - a_x^2} \end{aligned} \tag{3}$$

where the thermal excitation process F_x is a normally distributed random variable, i.e., $\mathcal{N}(0,1)$, and $\Delta T = 1/f_v$. The other model parameters suggested by [31] are $\beta_x = 10\ \text{s}^{-1}$, and $v_x = 100\ \text{cm/s}^{-1}$. The dynamics and parameters for the other Cartesian coordinates are the same.

The above state model which includes independent single speaker dynamics is formulated for the multispeaker state model as

$$p(\mathbf{X}_k|\mathbf{X}_{k-1}) \propto \prod_i p(\mathbf{X}_{i,k}|\mathbf{X}_{i,k-1}) \tag{4}$$

where $p(\mathbf{X}_{i,k}|\mathbf{X}_{i,k-1})$ denotes the dynamics for speaker i. It is highlighted that $p(\mathbf{X}_k|\mathbf{X}_{k-1})$ can be factorized for individual speakers.

B. Measurement Model

Visual measurements used in this work are based on the Viola-Jones face detector. The Viola-Jones face detector [51] yields good performance and detects faces extremely rapidly, by using a boosted cascade of features. It is a cascade of strong classifiers, each slightly more complex than the last. The input images are sub-sampled at multiple scales and locations to form the sub-windows for the faces to be detected. Face detection is performed in three stages. Initially, to minimize the effect of illuminations, the variance of all sub-windows are normalized. Second, the cascade of classifiers makes a decision based on the sub-window. Finally, to merge the overlapping face candidates around each face and output the final results the post processing method is used. A sub-window is detected as a face if it successfully passed by all strong classifiers. If any classifier fails a sub-window then no further processing is required on that window, detailed formulation is available in [51].

The center of the detected face is determined as the approximate position of the lips of the speaker in image coordinates $\boldsymbol{\iota}_{i,k}^c = [x_{i,k}, y_{i,k}]^T$, where c represents the number of cameras $c = 1, 2$. In 3-D space, each point in each camera frame defines a ray. Intersection of both rays is found by optimization methods, which finally help in calculation of the positions $\mathbf{Z}_{i,k}$ of the speakers in 3-D real world coordinates [56].

The multispeaker measurement model can be factorized in terms of individual speakers as

$$p(\mathbf{Z}_k|\mathbf{X}_k) = \prod_i p(\mathbf{Z}_{i,k}|\mathbf{X}_{i,k}) \tag{5}$$

where $p(\mathbf{Z}_{i,k}|\mathbf{X}_{i,k})$ is the observation model for speaker i and is calculated as

$$p(\mathbf{Z}_{i,k}|\mathbf{X}_{i,k}) \propto \exp\left(-\frac{\left\|\mathbf{Z}_{i,k} - \mathbf{X}'_{i,k}\right\|^2}{2\sigma^2}\right) \tag{6}$$

where $\mathbf{X}'_{i,k}$ denotes a vector formed from the 3-D position components of the state vector and σ is a standard deviation parameter chosen empirically, typically unity.

C. MCMC-Sampling Mechanism

In the 1990s, MCMC-based methods attracted great attention among researchers in the Bayesian community [57]. The advantage over alternative approaches is in the capacity to work with a high-dimensional space and complex models. It is computationally infeasible to track multiple objects in the high-dimensional space by using an importance sampling [52]-based traditional particle filter [58]. In tracking the MCMC sampling is a methodology for generating samples from a Markov chain whose stationary distribution corresponds to a filtering distribution. In order to efficiently place samples as close as possible to regions of high likelihood and approximate $p(\mathbf{X}_k|\mathbf{Z}_{1:k})$ in (2) with MCMC techniques, it is important to specifically design a Metropolis–Hastings (MH) sampler (also known as MCMC sampler) at each time step [40], [59], [60]. After running the MCMC sampler for long enough at each time step the initial part of the run, called the burn in period, is discarded to achieve a stationary distribution [61]. The key to the efficiency of the MCMC algorithm rests in the proposal distribution (discussed in the sequel), in which the configuration of one single object is modified at each step of the Markov chain, and each move in the chain is accepted or rejected by the acceptance ratio α. The MCMC-based tracking algorithm is summarized as follows:

- Initialize the MCMC sampler: At time k predict the state of each speaker i for N_p particles, i.e., $\{\mathbf{X}^n_{i,k}, \omega^n_k\}^{N_p}_{n=1}$ from the particle set at time $k-1$, i.e., $\{\mathbf{X}^n_{i,k-1}, \omega^n_{k-1}\}^{N_p}_{n=1}$ based on the factorized dynamic model $\prod_i p(\mathbf{X}_{i,k}|\mathbf{X}^n_{i,k-1})$.
- $B + N_p$ MCMC Sampling Steps: B and N_p denote the number of particles in the burn-in period and fair sample sets, respectively.
  - — Randomly select a speaker i from all speakers. This will be the speaker proposed to move.
  - — Sample a new state $\mathbf{X}^*_{i,k}$ for only speaker i from the single speaker proposal density $Q(\mathbf{X}^*_{i,k}|\mathbf{X}_{i,k})$.
  - — Compute the acceptance ratio which involves (2) for the evaluation of likelihood for only speaker i:

$$\alpha = min\left\{1, \frac{p\left(\mathbf{Z}_{i,k}|\mathbf{X}^*_{i,k}\right) Q\left(\mathbf{X}_{i,k}|\mathbf{X}^*_{i,k}\right)}{p(\mathbf{Z}_{i,k}|\mathbf{X}_{i,k}) Q\left(\mathbf{X}^*_{i,k}|\mathbf{X}_{i,k}\right)}\right\}. \quad (7)$$

  - — Draw $\mu \sim U(u|0,1)$.
  - — If $\alpha > \mu$ then accept the move for speaker i and change the $\mathbf{X}^*_{i,k}$ into $\mathbf{X}_k$. Otherwise, reject the move, do not change $\mathbf{X}_k$ and copy to the new sample set.
- Discard the first B samples to form the particle set, $\{\mathbf{X}^n_k, \omega^n_k\}^{N_p}_{n=1}$, at time step k.

The output of the 3-D tracker at each state k is the mean estimate for each speaker i and is calculated as the weighted sum over the associated particles as $\hat{\mathbf{x}}_{i,k} = \sum^{N_p}_{n=1} w^n_{i,k}\mathbf{x}^n_{i,k} / \sum^{N_p}_{n=1} w^n_{i,k}$, where in this work as in [40] $w^n_{i,k} = 1/N_p$.

1) Discussion on Algorithm Choice:

- State model (4) and measurement model (5) are independent in terms of the speakers and therefore can be factorized into the product of the marginal models for each speaker. In teleconferencing applications within an intelligent office, physical separation in speakers is always likely to be possible as the speakers are unlikely to embrace each other and speakers are also clearly separable in the 3-D real world coordinates used in this paper due in part to the height of the cameras. Therefore, in this work, there is no requirement to incorporate interaction cues in the state model. However, the state model could be extended with an interaction term in future work as in [59].
- Since joint particle filtering suffers from exponential complexity in the number of targets to be tracked [59], therefore independent particle filters for all speakers and an MCMC-PF are applicable for the requirement of the work in this paper. Independent SIR-PF for each speaker is the most optimal choice and is already used in our work [53]. Results show no significant difference (based on Euclidean error) but MCMC-PF reduces the computational complexity in multispeaker tracking.
- Due to the limitation of importance sampling in high-dimensional state space, MCMC methods are used. The MCMC method used in this work is based on [59], [60] and has the appealing property that *"the filter behaves as a set of individual particle filters when the targets are not interacting, but efficiently deals with complicated interactions when targets approach each other"*. The design of proposal density plays an important role in the success of an MCMC algorithm. The proposal density used is also defined in [40] as

$$Q\left(\mathbf{X}^*_k|\mathbf{X}_k\right) = \sum_{i^*} Q(i^*) Q\left(\mathbf{X}^*_k|\mathbf{X}_k, i^*\right) \quad (8)$$

  where a single speaker is first chosen with probability $Q(i^* = i)$ and a move is attempted on i (shown in the algorithm summary) and the rest of the multi-speaker configuration is left unchanged, where

$$Q\left(\mathbf{X}^*_k|\mathbf{X}_k, i^*\right) \propto \frac{1}{N_p}\sum_n p\left(\mathbf{X}^*_{i^*,k}|\mathbf{X}^n_{i^*,k-1}\right) \times \prod_{l\in m-\{i^*\}} \delta\left(\mathbf{X}^*_{l,k} - \mathbf{X}_{l,k}\right) \quad (9)$$

  and $\mathbf{X}_k$ represents the whole state for all speakers m, $\delta(.)$ is a delta function, and $\mathbf{X}_{i^*,k}$ denotes the substate for one speaker. It is highlighted that the proposal density used in [40] appears not to be properly formulated. This has thus been modified in (9).

The change in the position of a speaker with respect to the previous state $k-1$ (known as velocity information) also plays a critical role to decide the method for source separation and is discussed next.

V. 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 this visual information, we decide either the sources should be separated as moving or stationary.

The pseudo code to issue the command for selecting the source separation methods are as follows.

Pseudo Code: Command for Selecting the Source Separation Methods

- Reset the $counter$ and set the $threshold$
- FOR $j = 2 : k$
  -Find $d_{i,j} = \|\hat{\mathbf{X}}_{i,j} - \hat{\mathbf{X}}_{i,j-1}\|_2$
- IF $d_{i,j} < threshold$
  -Update the $counter$
  $*$IF $counter > T_k/T_v$
  -Command for the FastICA based method
  $*$END IF
  -Set $\hat{\mathbf{X}}_{i,j} = \hat{\mathbf{X}}_{i,j-1}$
  ELSE
  -Command for the beamforming based method
  -Reset the $counter$
- END IF
- END FOR
  THIS CODE WOULD BE USED FOR EACH SPEAKER i.

where T_k represents the expected stationary period for the speakers, $T_v = 1/f_v$, $\|.\|_2$ denotes Euclidean norm, and $threshold$ is the minimum distance required for the beamformer channel filters, which should be recalculated to separate the sources.

When the sources are physically stationary for a certain period T_k we separate the sources with IIFastICA. By changing the value of T_k we can change the expected required stationary period for the sources.

The other important parameter to be calculated before starting the source separation is the angle of arrival of each speaker to the sensor array as shown in Fig. 2. By having the position information of the microphones and the speakers at each state k from the 3-D visual tracker, we can easily calculate the angle of arrival $\boldsymbol{\theta}_{i,k}$ of speakers relative to the microphone sensor array.

With $\boldsymbol{\theta}_{i,k}$ and the control command from the above decision criterion at each state k, we separate the sources either by beamforming or by IIFastICA as discussed in the following subsections.

A. Beamforming-Based Separation

In the intelligent office where our recordings are taken, the microphones used are unidirectional. By using a short-time discrete Fourier transform (DFT) the mixing process can be formulated as follows: having M statistically independent real sources $\mathbf{s}(\omega) = [s_1(\omega), \ldots, s_M(\omega)]^H$ where ω denotes discrete normalized frequency, a multichannel finite impulse response (FIR) filter $\mathbf{H}(\omega)$ producing N observed mixed signals $\mathbf{u}(\omega) = [u_1(\omega), \ldots, u_N(\omega)]^H$, where $(.)^H$ is Hermitian transpose, can be described as (we assume there is no noise or noise can be deemed as a source signal in the model for simplicity)

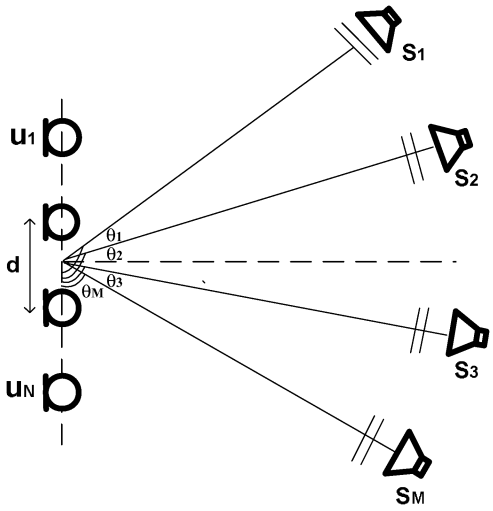

Fig. 2. Microphone and source layout.

$$\mathbf{u}(\omega) = \mathbf{H}(\omega)\mathbf{s}(\omega) \tag{10}$$

where

$$\mathbf{H}(\omega) = \begin{bmatrix} h_{11}(\omega) \cdots h_{1M}(\omega) \\ \vdots \quad \vdots \\ h_{N1}(\omega) \cdots h_{NM}(\omega) \end{bmatrix} \tag{11}$$

and the source separation can be described as

$$\mathbf{y}(\omega) = \mathbf{W}(\omega)\mathbf{u}(\omega) \tag{12}$$

where

$$\mathbf{W}(\omega) = \begin{bmatrix} w_{11}(\omega) \cdots w_{1N}(\omega) \\ \vdots \quad \vdots \\ w_{M1}(\omega) \cdots w_{MN}(\omega) \end{bmatrix}. \tag{13}$$

$\mathbf{y}(\omega) = [y_1(\omega), \ldots, y_N(\omega)]^H$ contains the estimated sources, and $\mathbf{W}(w)$ is the unmixing filter matrix. An inverse short time Fourier transform is then used to find the estimated sources $\hat{\mathbf{s}}(t) = \mathbf{y}(t)$. In this work to demonstrate the proposed approach we consider the exactly determined convolutive BSS problem, i.e., $N = M = 2$.

The unmixing matrix $\mathbf{W}(\omega)$ for each frequency bin can be approximated from beamforming methods. In recent years, many beamforming methods have been proposed, such as the linearly constrained minimum variance (LCMV) method and the minimum variance distortionless response (MVDR) method [62]. A post filtering approach has also been utilized to improve these methods [63]. However, the LCMV method and the MVDR method need estimates of statistical information of the input or noise signals, which are not accessible in the context of BSS for moving sources. Furthermore, a diffuse noise field assumption is used in [63], which is not valid in the context of BSS. The postfilter method has been used in [47] to perform speech enhancement for both stationary and moving cases; however,

the model used in that speech enhancement is a single-input single-output (SISO) model, which is different from the multi-input multi-output (MIMO) model in the context of BSS; thus, this postfiltering approach is also not suitable for the solution of BSS. To the best of our knowledge, in the context of BSS, only the beamforming approach that is directly obtained from the inverse of the mixing matrix model has been successfully used in [64]. To compensate for noninvertibility of the mixing matrix, a regularization term is included in [65], in which the beam pattern obtained from the geometric information is incorporated in the solution of BSS. Similar to that in [65], the unmixing matrix in our approach is calculated as

$$\mathbf{W}(\omega) = \left(\mathbf{H}(\omega)^H \mathbf{H}(\omega) + \beta I\right)^{-1} \mathbf{H}(\omega)^H \tag{14}$$

where $\mathbf{W}(\omega) = [\mathbf{w}_1(\omega), \ldots, \mathbf{w}_N(\omega)]$, $\mathbf{H}(\omega) = [\mathbf{h}_1(\omega), \ldots, \mathbf{h}_M(\omega)]$, β is a small positive constant such as 0.01 in our simulations, and I represents the identity matrix.

The delay element between source l and sensor k, i.e., $h_{kl}(\omega)$ is calculated as

$$h_{kl}(\omega) = e^{j(k-l)d\cos(\theta_l)\omega/c} \quad k = 1, \ldots, N \quad l = 1, \ldots, M \tag{15}$$

where d is the distance between the sensors and c is the speed of sound in air.

Ideally, $h_{kl}(\omega)$ should be the sum of all echo paths, but these cannot all be found; therefore, an approximation is used by neglecting the room reverberations.

Finally, by placing $\mathbf{W}(\omega)$ in (12), we estimate the sources. Since the scaling is not a major issue [6] and there is no permutation problem, we can therefore align the estimated sources for reconstruction in the time domain.

B. FastICA Based Separation

If the sources are stationary for at least two seconds, we extract the sources with the help of the estimated $\mathbf{H}(\omega)$ from the above section and the whitening matrix for the mixtures, as an initialization of the FastICA algorithm [66]. We thereby improve the convergence of the algorithm and also increase the separation performance together with mitigate the permutation problem. Crucially, in the frequency domain convolutive BSS (FDCBSS) approach, since the algorithm essentially fixes the permutation at each frequency bin, there will be no problem while aligning the estimated sources for reconstruction in the time domain.

As an initial step, it is usual in ICA approaches to sphere or whiten the data

$$\mathbf{z}(\omega) = \mathbf{Q}(\boldsymbol{\omega})\mathbf{u}(\boldsymbol{\omega}) \tag{16}$$

where $\mathbf{Q}(\omega)$ is the whitening matrix [67].

Each column of $\mathbf{H}(\omega)$ is used to initialize the fixed point algorithm [66] for each frequency bin:

$$\mathbf{w}_i(\omega) = \mathbf{Q}(\omega)\mathbf{h}_i(\omega). \tag{17}$$

We have the following approximate Newton iteration for each vector of each frequency bin [66]

$$\begin{aligned}\mathbf{w}_i^+(\omega) = {} & E\left\{\mathbf{z}(\omega)\left(\mathbf{w}_i(\omega)^H\mathbf{z}(\omega)\right)^* g\left(\left|\mathbf{w}_i(\omega)^H\mathbf{z}(\omega)\right|^2\right)\right\} \\ & - E\left\{g\left(\left|\mathbf{w}_i(\omega)^H\mathbf{z}(\omega)\right|^2\right) + \left|\mathbf{w}_i(\omega)^H\mathbf{z}(\omega)\right|^2 \right. \\ & \left. \times \acute{g}\left(\left|\mathbf{w}_i(\omega)^H\mathbf{z}(\omega)\right|^2\right)\right\}\mathbf{w}_i(\omega) \\ \mathbf{w}_i(\omega) = {} & \frac{\mathbf{w}_i^+(\omega)}{\left\|\mathbf{w}_i^+(\omega)\right\|}\end{aligned} \tag{18}$$

where $(.)^*$ denotes the complex conjugate, $g(.)$ and $\acute{g}(.)$ denote the first and second derivative of the contrast function $G(|\mathbf{w}^H(\omega)\mathbf{z}(\omega)|^2)$. In the experiments, the statistical expectation is realized as a sample average.

We have N independent components, i.e., $\mathbf{w}_i(\omega)$, $i = 2, \cdots, N$, which are calculated in parallel to obtain $\mathbf{W}(\omega) = [\mathbf{w}_1(\omega), \cdots, \mathbf{w}_N(\omega)]$ for each frequency bin. After each iteration the independent components are decorrelated in a symmetric orthogonalization scheme which is more accurate than a deflationary orthogonalization in the exactly determined case we are addressing. The symmetric orthogonalization takes the form [67]

$$\mathbf{W}(\omega) = \mathbf{W}(\omega)\left\{\mathbf{W}^H(\omega)\mathbf{W}(\omega)\right\}^{-\frac{1}{2}}. \tag{19}$$

Before starting the update process, $\mathbf{H}(\omega)$ is normalized once using $\mathbf{H}(\omega) \leftarrow \mathbf{H}(\omega)/\|\mathbf{H}(\omega)\|_F$, where $\|.\|_F$ denotes the Frobenius norm. Finally, by placing $\mathbf{W}(\omega)$ in (12) we recover the sources. Next, we evaluate the proposed schemes by simulation studies.

VI. Experiments and Results

A. Setup and Evaluation Criterion

1) Data Collection: The simulations are performed on real recorded audio-visual signals generated from a room geometry as illustrated in Fig. 3. Data are collected in a $4.6 \times 3.5 \times 2.5$ m^3 intelligent office. Two out of eight calibrated color video cameras ($C1$ and $C2$ shown in Fig. 3) are utilized to collect the video data. Video cameras are fully synchronized with an external hardware trigger module and frames are captured at $f_v = 25$ Hz with an image size of 640 × 480 pixels, frames were down-scaled if it was necessary, and reducing the resolution by half was a good tradeoff between accuracy and resolution. Both video cameras have overlapping fields of view. The duration between consecutive states is $T_v = 1/25$ s. Audio recordings are taken at $f_a = 8$ kHz and are synchronized manually with video recordings. The distance between the audio-sensors is $d = 4$ cm. The other important variables are selected as: number of sensors and speakers $N = M = 2$, number of particles $N_p = 600$, $B = 200$, the number of images in the first and second experiment are $k = 525$ and 600, which respectively indicate 21 and 24 seconds of data, $T_k = 5$ s, $threshold = 0.04$ m, speed of sound in air $c = 343$ m/s, FFT length $T = 2048$ and filter length $Q = 1024$, height of the cameras in the intelligent office is 2.35 m, and the room impulse duration is 130 ms. In the proposed algorithm the nonlinearity for FastICA is $G(y) = \log(b + y)$,

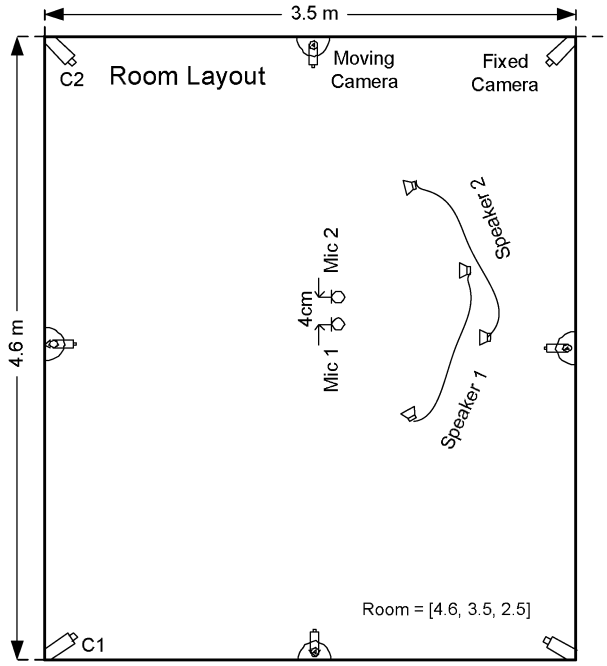

Fig. 3. Two-speaker two-microphone layout for recording within a reverberant (room) environment. Room impulse response length is 130 ms.

with $b = 0.1$. In the first experiment on tracking, speaker 2 is stationary and speaker 1 is moving and in the second experiment both speakers are moving around a table in a teleconference scenario.

2) BSS Evaluation Criterion: In this paper, the performance of the algorithm is evaluated on the basis of two criteria on real room recordings. The signal-to-interference ratio (SIR) is calculated as in [6]

$$SIR = \frac{\Sigma_i \Sigma_\omega |H_{ii}(\omega)|^2 \left\langle |s_i(\omega)|^2 \right\rangle}{\Sigma_i \Sigma_{i \neq j} \Sigma_\omega |H_{ij}(\omega)|^2 \left\langle |s_j(\omega)|^2 \right\rangle} \tag{20}$$

where H_{ii} and H_{ij} represent, respectively, the diagonal and off-diagonal elements of the frequency domain mixing filter, and s_i is the frequency domain representation of the source of interest.

Second, in order to evaluate the source separation with solution to permutation by integrating audio-visual information in the initialization of FastICA, we use the Performance Index (PI) measurement which provides results at each frequency bin level. The PI as a function of the overall system matrix $\mathbf{G} = \mathbf{WH}$ is given as

$$PI(G) = \left[\frac{1}{n} \sum_{i=1}^{n} \left(\sum_{k=1}^{m} \frac{abs(G_{ik})}{max_k abs(G_{ik})} - 1 \right) \right] + \left[\frac{1}{m} \sum_{k=1}^{m} \left(\sum_{i=1}^{n} \frac{abs(G_{ik})}{max_i abs(G_{ik})} - 1 \right) \right] \tag{21}$$

where G_{ik} is the ikth element of $\mathbf{G}$.

As we know the above PI based on [3] is insensitive to permutation. We therefore evaluated the permutation on the basis of the criterion, i.e., $[abs(G_{11}G_{22}) - abs(G_{12}G_{21})] > 0$ for permutation free FDCBSS, used in our works [8], [53].

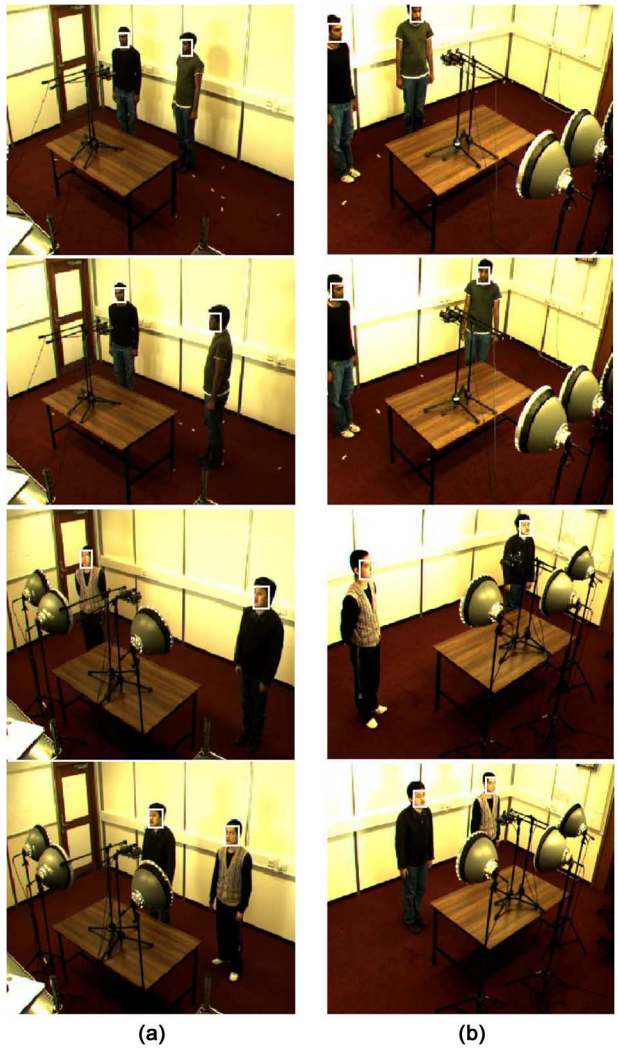

Fig. 4. 3-D tracking results 1: frames of synchronized recordings, (a) frames of first camera, and (b) frames of second camera; the Viola-Jones face detector [51] efficiently detected the faces in the frames.

B. Results and Discussion

1) 3-D Tracking Results: In this section, the results obtained from tracking are discussed. Two experiments are performed to evaluate the 3-D visual tracker. The faces of the speakers are detected by using the Viola-Jones face detector [51] which efficiently detected the faces in the frames shown in Fig. 4. Since in the dense environment as shown in Fig. 4 it is very hard to detect the lips directly, therefore the center of the detected face region as the position of the lips in each sequence is approximated. More sophisticated and computationally efficient schemes could also be proposed for detecting the face through a sequence of images but the approach adopted in this work is sufficient to verify the multimodal CBSS method, the target of this work.

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 by using the calibration parameters [46] and the optimization method [56]. With this measurement the particle filter is updated. The number of particles in both experiments for MCMC-PF was the same $N_p = 600$, $B = 200$, for SIR-PF was $N_p = 1000$ and results were obtained using single runs. These parameters have been determined empirically

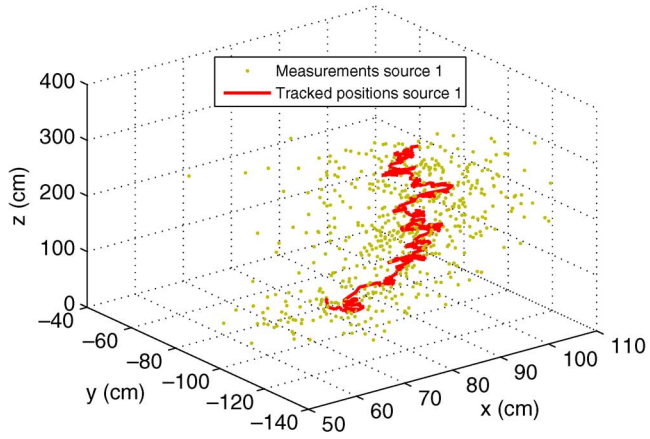

Fig. 5. 3-D tracking results 1: SIR-PF-based 3-D tracking of speaker 1 while walking around the table in the intelligent office. Speaker 2 is physically stationary in this experiment.

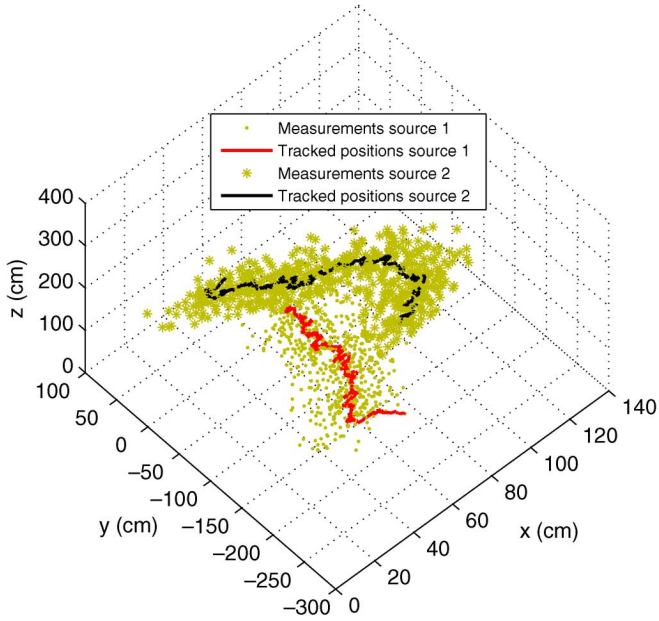

Fig. 6. 3-D tracking results 2: SIR-PF-based 3-D tracking of the speakers while walking around the table in the intelligent office.

to provide a good compromise between algorithm performance and computational complexity.

In the first experiment, speaker 2 is stationary and speaker 1 is moving around the table so the tracking results of the speaker 1 are discussed in detail in this experiment. The sampling importance resampling particle filter (SIR-PF) is also suitable for this case as used in our work [53]. In the second experiment, both speakers are simultaneously moving and their motion is more complicated as they cross over. MCMC-PF is suitable for multispeaker tracking because it improves the sampling efficiency with approximately the same computational cost of the Generic-PF. In the second experiment, both SIR-PF and MCMC-PF are used. The gait of the speakers is not smooth and the speakers are also stationary for a while at some points during walking around the table which provides a good test for the evaluation of the 3-D tracker as well as for source separation methods, and this is also clear in the 3-D tracking results shown in Figs. 5–7.

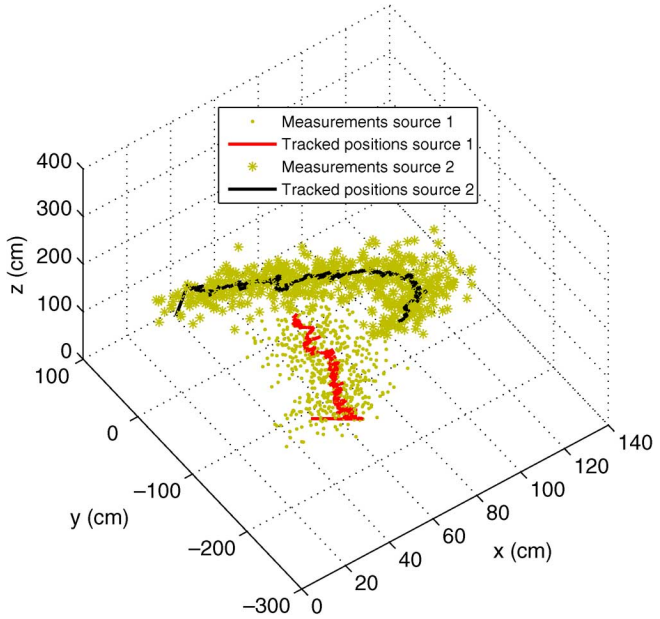

Fig. 7. 3-D tracking results 2: MCMC-PF-based 3-D tracking of the speakers while walking around the table in the intelligent office.

In order to view the tracking results in more detail, the tracking results are plotted in the x, y, and z axes separately. Figs. 8–10 clearly show that tracking result has removed much of the measurement uncertainty and later in this section the error in detection for the particle filter will be quantified. The benefit of the true 3-D tracker is clearly shown in Fig. 10. In particular, although the speakers would approximately coalesce in the 2-D image plane, they are clearly separable in the 3-D real world coordinates due in part to the height of the cameras. In 2-D tracking in the image plane this problem cannot be avoided. The error bars for particle filters at different states are also plotted in these results. It is highlighted that the error bars would appear as 3-D surfaces in the pseudo-3-D plots and would make the plots cluttered if they were displayed. However, the behavior of the error ellipses on the 2-D plots gives a clear indication as to how 3-D error bar surfaces would appear on the 3-D plots.

Actually, the height of the speakers is fixed and during walking only the movement in the heads will produce minor change which is clear in Figs. 11–13. Since the speakers and microphones are approximately at the same level, therefore it is assumed that effective movement is in the xy plane.

In order to evaluate the performance of the tracker as in [47], the Euclidean distance to the frame-based ground truth is generated at each state. To calculate the ground truth, a time-consuming manual task is performed by annotating the mouth position of each speaker in each camera frame. Figs. 14–16 provide the Euclidean error at each state for both experiments. In the first experiment, the mean error is 0.05 m and standard deviation is 0.03 m. In the second experiment, the mean error is 0.055 m and standard deviation is 0.032 m which confirm the good performance of the tracker.

2) Angle of Arrival Results: The calculated position of the center of the microphones in experiment 1 is $[-0.08, -0.22, 1.62]^T$ m, the position of speaker 2 is

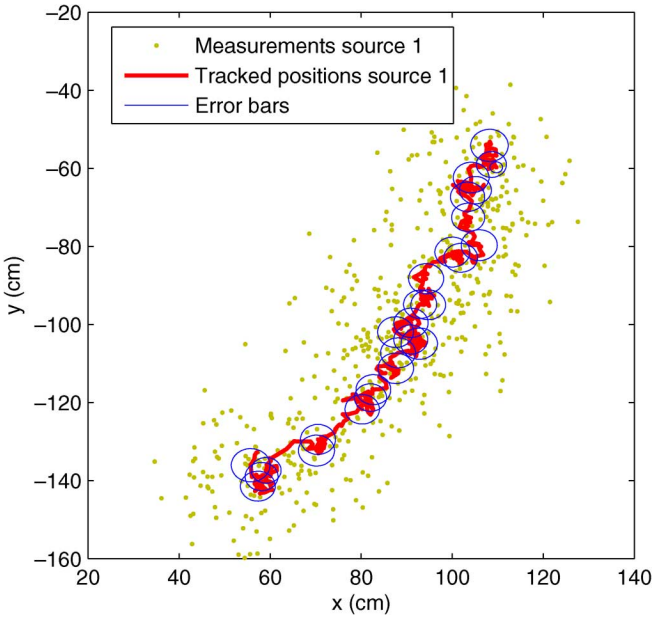

Fig. 8. 3-D tracking results 1: SIR-PF-based tracking of the speaker 1 in the x and y axis, while walking around the table in the intelligent office. Speaker 2 is physically stationary in this experiment. The result provides more in depth view in the x and$y$ axis.

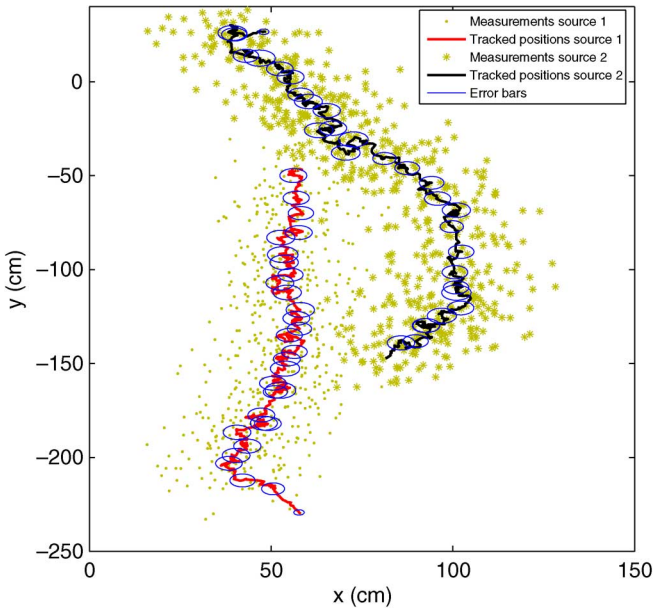

Fig. 9. 3-D tracking results 2: SIR-PF-based tracking of the speakers in the x and y axis, while walking around the table in the intelligent office. The result provides more in depth view in the x and y axis.

$[0.94, 0.59, 1.63]^T$ m (the reference point in the room is under the table, close to the microphones) and the tracked position of speaker 1 in states $k = 1 : 525$ is shown in Fig. 8. The angle of arrival of speaker 2 is 128° and the angles of arrivals of the speaker 1 are shown in Fig. 17. The calculated position of the center of the microphones in experiment 2 is $[-0.50, -0.94, 1.60]^T$ m. The angle of arrivals of both speakers are shown in Fig. 18. In the results of both experiments, it is found that the effective movement of the speakers were in the x axis and y axis; therefore, the effective change in the angle of arrival was only in the xy plane.

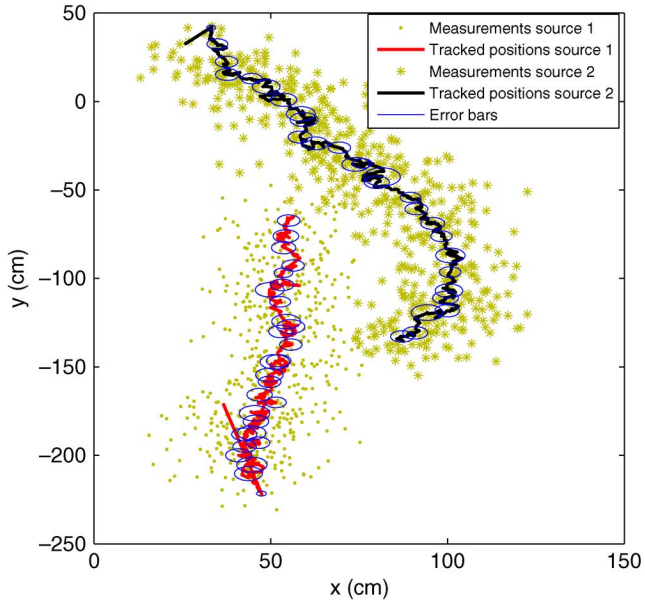

Fig. 10. 3-D tracking results 2: MCMC-PF-based tracking of the speakers in the x and y-axis, while walking around the table in the intelligent office. The result provides more in depth view in the x and y axis.

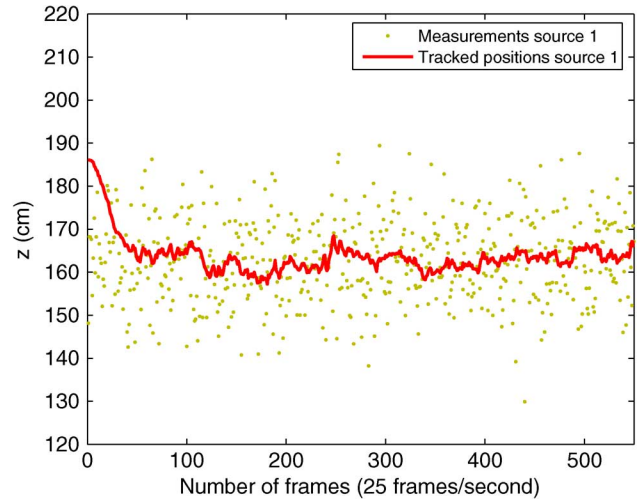

Fig. 11. 3-D tracking results 1: SIR-PF-based tracking of the speaker 1 in the z axis, while walking around the table in the intelligent office. Speaker 2 is physically stationary in this experiment. The result confirms that there is very small change in the z axis with respect to the x and y axis.

Now a successful tracker is available to provide the required geometric information to perform multimodal blind source separation. Therefore, simulations on BSS are discussed next.

3) BSS Results: The objective evaluation of BSS is limited by the requirement of the mixing filter; therefore, for such testing the audio signals are convolved with real room impulse responses recorded in certain positions of the room. The separation of the real recorded signals in the intelligent office is evaluated subjectively by listening tests and mean opinion scores (MOSs) are provided at the end. In the context of objective evaluation, termed moving source test (MST), it is assumed that the moving sources remain static over a particular time interval less than 0.5 s. The justification is that over this interval no frequency domain CBSS algorithm could be used as there would be insufficient number of samples to achieve

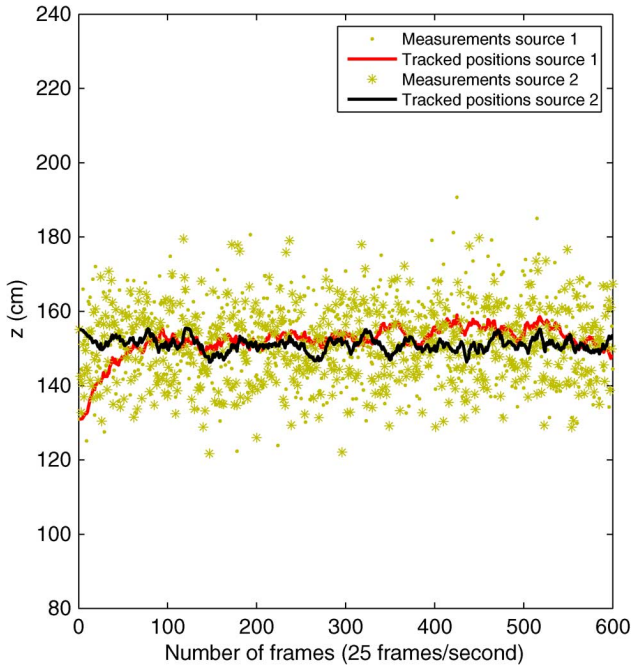

Fig. 12. 3-D tracking results 2: SIR-PF based tracking of the speakers in the z axis, while walking around the table in the intelligent office. The result confirms that there is very small change in the z axis with respect to the x and y axis.

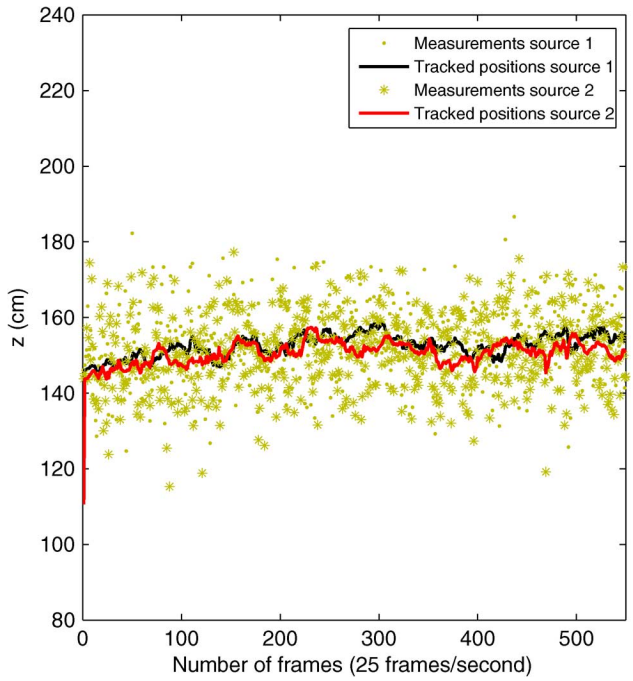

Fig. 13. 3-D tracking results 2: MCMC-PF-based tracking of the speakers in the z axis, while walking around the table in the intelligent office. The result confirms that there is very small change in the z axis with respect to the x and y axis.

convergence, but the proposed beamforming is successful as it is independent of data length.

Five simulations for comparison of the proposed algorithm are presented.

- FastICA [66] (Matlab code available online)-based BSS with random initialization and length of the signals is 5 s.
- FastICA based BSS with intelligent initialization and length of the signals is 5 s.

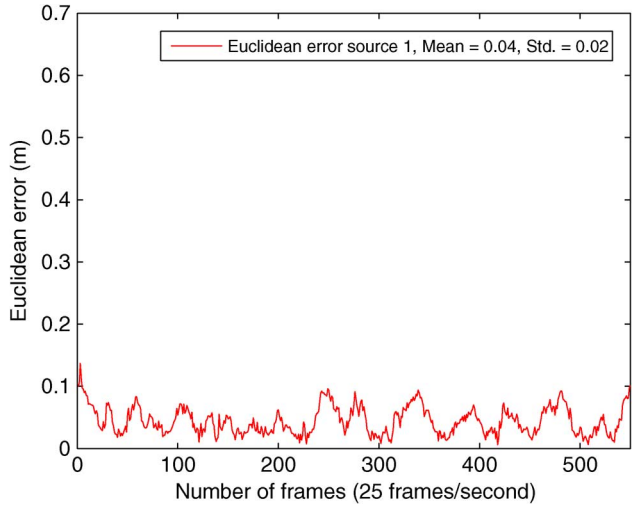

Fig. 14. 3-D tracking results 1: SIR-PF-based tracking of the speaker 1. Speaker 2 is physically stationary. Euclidean error is calculated against manually annotated frame-based ground truths in each camera plane of speaker 1.

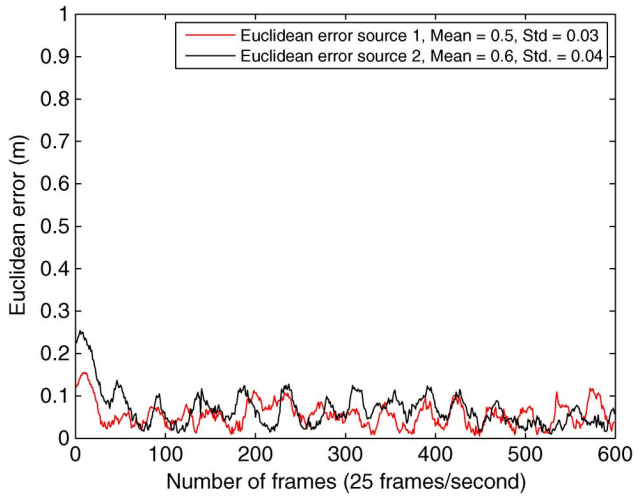

Fig. 15. 3-D tracking results 2: SIR-PF-based tracking of the speakers. Euclidean error is calculated against manually annotated frame-based ground truths in each camera plane of the speakers.

- FastICA based BSS with intelligent initialization but length of the signals is 0.4 s (the MST case).
- Beamforming based BSS and the length of the signals is 0.4 s (the MST case).
- Beamforming based BSS when both sources are physically close to each other.

Initially, in the first simulation the recorded mixtures of length of 5 s are separated by the original FastICA algorithm. The performance indices and evaluation of permutation by the original FastICA algorithm [66] with random initialization are shown in Fig. 19. It is highlighted that 35 iterations are required for the performance level achieved in Fig. 19(a) with no solution for permutation as shown in Fig. 19(b). The permutation problem in frequency domain BSS degrades significantly the separation performance for the recorded mixtures.

In the second simulation, recorded mixtures of length of 5 s are again separated. In this simulation, the angle of arrival of both speakers obtained from the 3-D tracker is passed one-by-one to (15) and FastICA is intelligently initialized (as

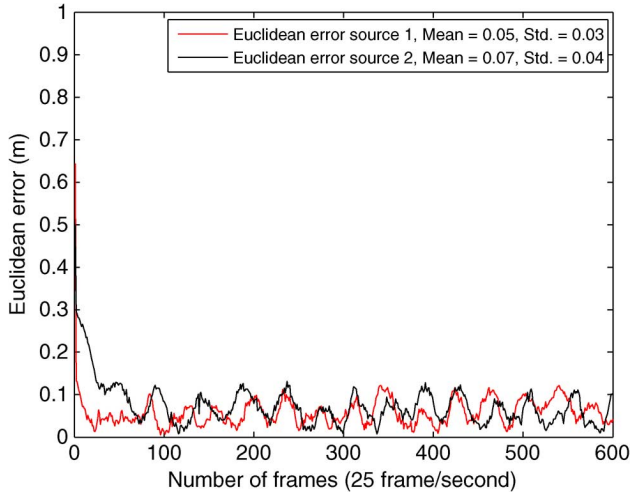

Fig. 16. 3-D Tracking results 2: MCMC-PF-based tracking of the speakers. Euclidean error is calculated against manually annotated frame-based ground truths in each camera plane of the speakers.

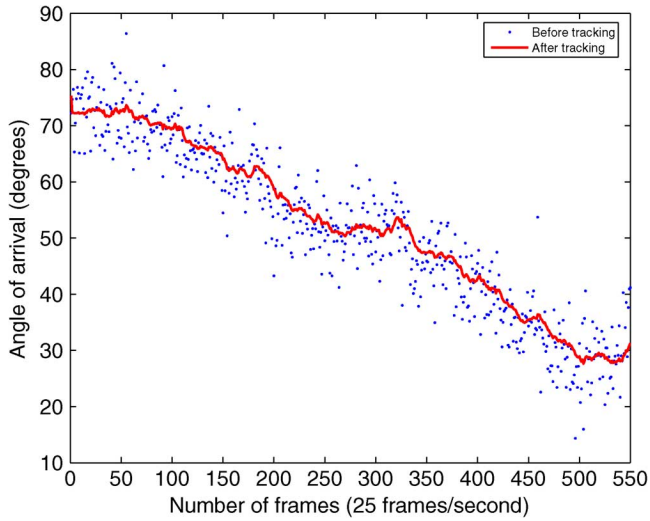

Fig. 17. Angle of arrival results 1: Angle of arrival of speaker 1 relative to the sensor array. Speaker 2 is physically stationary in this experiment. The estimated angle before tracking and corrected angle by SIR-PF are shown. The change in angle is not smooth because of the gait of the speaker.

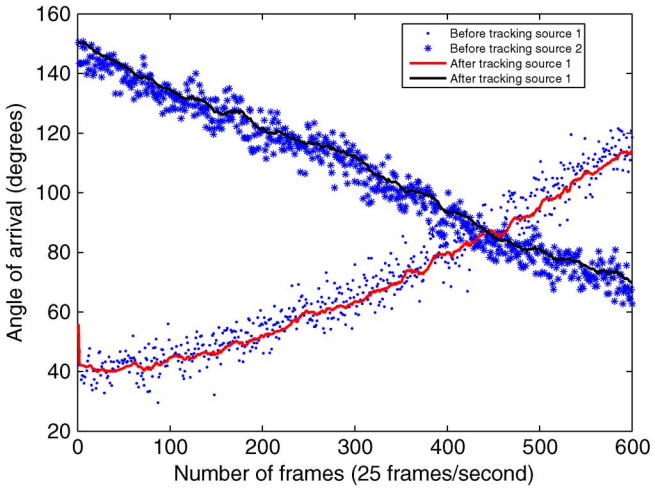

Fig. 18. Angle of arrival results: Angle of arrival of the speakers to the sensor array. The estimated angle before tracking and corrected angle by MCMC-PF are shown.

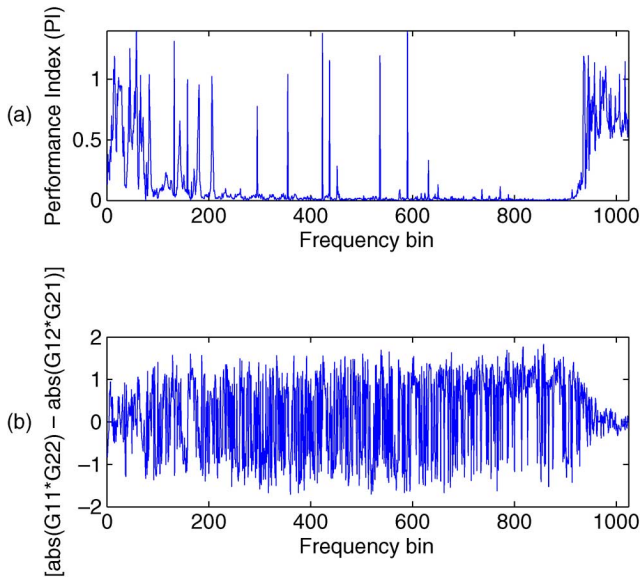

Fig. 19. BSS Results: performance index at each frequency bin for the original Bingham and Hyvärinen algorithm on the top [66] and evaluation of permutation at the bottom, on the recorded signals of known room impulse response with fixed iteration count $= 35$, length of the signals is 5 s. A lower PI refers to a superior method and $[abs(G_{11}G_{22}) - abs(G_{12}G_{21})] > 0$ means no permutation. Audio sampling frequency $f_a = 8$ kHz and FFT length $T = 2048$.

discussed in Section V-B). The resulting performance indices are shown in Fig. 20(a) which shows good performance, i.e., close to zero across the majority of the frequency bins. This is due to visual information used in the initialization, and the algorithm also converges in six iterations. The visual modality therefore renders this BSS algorithm semiblind and thereby much improves the resulting performance and the rate of convergence. Permutation is evaluated on the basis of the criterion mentioned above. In Fig. 20(b), the results confirm that the proposed algorithm automatically mitigates the permutation at each frequency bin. Since there is no permutation problem, the sources are therefore finally aligned in the time domain. In Fig. 20(a), at higher frequency bins there is less energy in the mixtures therefore performance in those bins is deteriorated. The SIR is also calculated as in [6] and results are shown in Table I.

In the third simulation, the length of the mixtures is reduced to 0.4 s, i.e., the MST case, and the performance is shown in Fig. 21. It is obvious in the results that the performance is poor because FastICA is based on fourth-order statistics and is limited by the data length requirement. For signals with length equal to 0.4 s, given the block length of the FFT, only one sample would be available at each frequency bin $round(0.4f_a/T) = 1$ and therefore batch-wise BSS algorithms cannot separate the sources of short data length due to insufficient samples to converge, which is a common problem when the sources are moving.

In the fourth simulation, the angles of arrival of both speakers obtained from the 3-D tracker are passed to (15) and the sources were separated by using beamforming (discussed in Section V-A) and the results are shown in Fig. 22. The resulting performance indices are shown in Fig. 22(a) and confirm good performance and Fig. 22(b) also shows that the

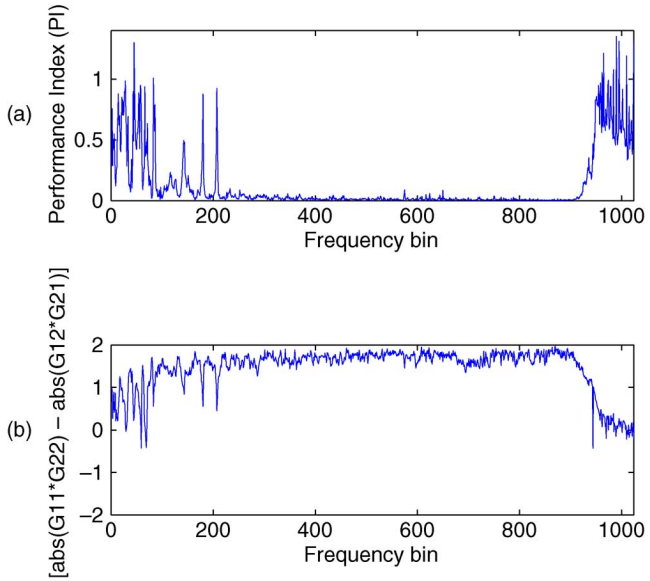

Fig. 20. BSS Results: performance index at each frequency bin for the proposed intelligently initialized FastICA algorithm at the top and evaluation of permutation at the bottom, on the recorded signals of known room impulse response with fixed iteration count $= 6$, length of the signals is 5 s. A lower PI refers to a superior method and $[abs(G_{11}G_{22}) - abs(G_{12}G_{21})] > 0$ means no permutation. Audio sampling frequency $f_a = 8$ kHz and FFT length $T = 2048$.

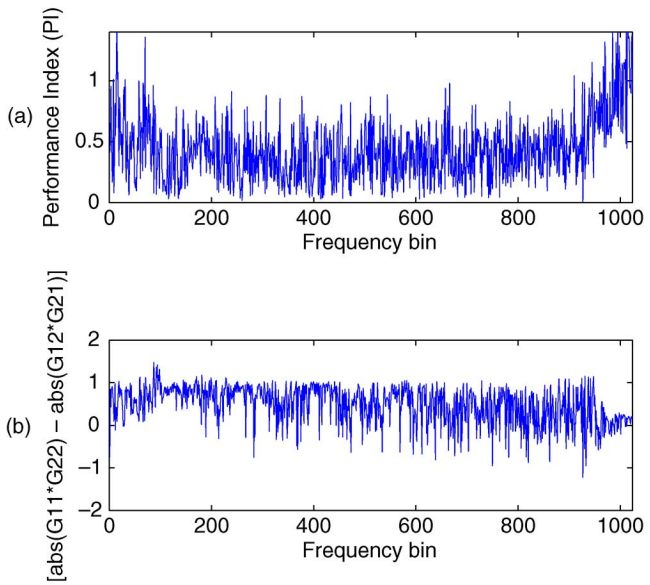

Fig. 21. BSS Results: performance index at each frequency bin for the proposed intelligently initialized FastICA algorithm at the top and evaluation of permutation at the bottom, on the recorded signals of known room impulse response, length of the signals is 0.4 s. A lower PI refers to a superior method and $[abs(G_{11}G_{22}) - abs(G_{12}G_{21})] > 0$ means no permutation. Audio sampling frequency $f_a = 8$ kHz and FFT length $T = 2048$.

TABLE I
BSS RESULTS: COMPARISON OF SIR-IMPROVEMENT BETWEEN ALGORITHMS AND THE PROPOSED METHOD FOR DIFFERENT SETS OF MIXTURES

Algorithms	SIR-Improvement (dB)
Parra et al. Method [5]	6.8
Wenwu et al. Method [4]	10.0
Hiroshi et al. Method [64]	11.7
IIFastICA	12.9

beamforming mitigates the permutation. Since there is no permutation problem, therefore the sources can be aligned in the time domain. For comparison the data length of the mixtures used in this simulation is 0.4 s and SIR in this case is 9.5 dB. It is know that the ideal condition for beamforming is when there is no reverberation in the room (instantaneous case), but is not possible in a real environment; however, the beamformer still works in a moderate reverberant environment as in this case (room impulse response length is 130 ms).

In the last simulation, when both speakers are physically close to each other, i.e., at state $k = 393$ (where both speakers are close and stationary for 0.4 s) the position of the speaker 1 is $[0.54, -1.10, 1.59]^T$ m and the position of speaker 2 is $[1.00, -0.91, 1.58]^T$ m, the angles of arrivals of both speakers, i.e., 81° and 91°, respectively, obtained from the above estimated positions from the 3-D tracker are passed to (15) and the sources are separated by using beamforming and the results are shown in Fig. 23. In this case, the performance reduces because of the limitations of the beamformer, i.e., it is unable to discriminate spatially one speaker from another due to the width of its mainlobe being greater than the separation of the speakers, which is particularly clear at lower frequencies. For comparison, the data length of the mixtures used in this simulation is also 0.4 s and SIR in this case is 8.2 dB. In conclusion,

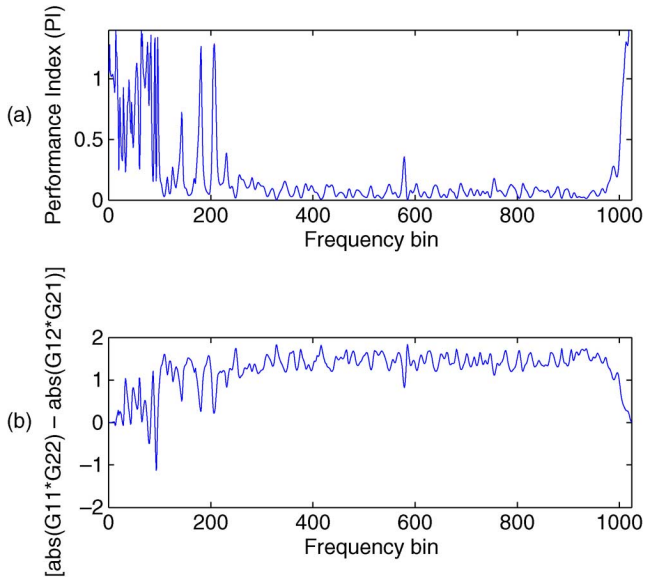

Fig. 22. BSS Results: performance index at each frequency bin for 3-D tracking based angle of arrival information used in beamforming at the top and evaluation of permutation at the bottom, on the recorded signals of known room impulse response, beamforming-based separation is independent of length of the signals. A lower PI refers to a superior method and $[abs(G_{11}G_{22}) - abs(G_{12}G_{21})] > 0$ means no permutation. Audio sampling frequency $f_a = 8$ kHz and FFT length $T = 2048$.

beamforming provides the solution for source separation of moving sources at an acceptable level because beamforming is independent of the data length requirement unlike second- or fourth-order statistics-based batch-wise BSS algorithms. 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 III

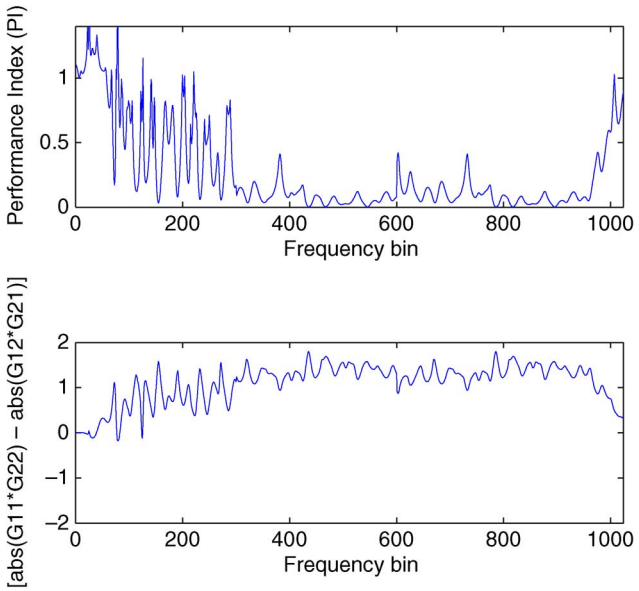

Fig. 23. BSS Results: performance index at each frequency bin for 3-D tracking based angle of arrival information used in beamforming at the top and evaluation of permutation at the bottom, on the recorded signals of known room impulse response, beamforming-based separation is independent of length of the signals. Speakers are physically close to each other therefore performance is reduced. A lower PI refers to a superior method and $[abs(G_{11}G_{22}) - abs(G_{12}G_{21})] > 0$ means no permutation. Audio sampling frequency $f_a = 8$ kHz and FFT length $T = 2048$.

TABLE II
LISTENING-QUALITY SCALE

Quality of the speech	Mean opinion score
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

TABLE III
SUBJECTIVE EVALUATION: MOS FOR SEPARATION OF REAL ROOM RECORDINGS, BY IIFastICA AND DIFFERENT ALGORITHMS WHEN SOURCES ARE PHYSICALLY STATIONARY, AND BY BEAMFORMING WHEN SOURCES ARE PHYSICALLY MOVING

Algorithms	Mean opinion score
Parra et al. Method [5]	4.1
Wenwu et al. Method [4]	4.3
Hiroshi et al. Method [64]	4.4
IIFastICA	4.7
Beamforming	3.8

(MOS tests for voice are specified by ITU-T recommendation P.800 and listening-quality scale is shown in Table II).

VII. CONCLUSION

In this paper, a new multimodal BSS approach is proposed to solve the moving source separation problem. A full 3-D tracker based on MCMC-PF is implemented. Video information is utilized in the 3-D tracker which provides velocity and direction information of sources. Based on the velocity of the source, a criterion for source separation is setup: a beamforming algorithm is used when sources are moving and a BSS algorithm is performed when sources are stationary. 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 an important step forward towards the solution of the real cocktail party problem. Further evaluations with, complex motions, multiple speakers, and postprocessing will be a future work.

ACKNOWLEDGMENT

The authors would like to thank Dr. Y. Zhang, Dr. S. Sanei, Dr. J. Hicks, and Dr. A. Aubrey for discussions and support to collect the data.

REFERENCES

[1] S. Haykin *et al., Unsupervised Adaptive Filtering (Volume I: Blind Source Separation)*. New York: Wiley, 2000.
[2] A. S. Bregman, *Auditory Scence Analysis*. Cambridge, MA: MIT Press, 1990.
[3] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*. New York: Wiley, 2002.
[4] W. Wang, S. Sanei, and J. Chambers, "Penalty function based joint diagonalization approach for convolutive blind separation of nonstationary sources," *IEEE Trans. Signal Process.*, vol. 53, no. 5, pp. 1654–1669, May 2005.
[5] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 320–327, Mar. 2000.
[6] S. Sanei, S. M. Naqvi, J. A. Chambers, and Y. Hicks, "A geometrically constrained multimodal approach for convolutive blind source separation," in *Proc. IEEE ICASSP*, 2007, pp. 969–972.
[7] T. Tsalaile, S. M. Naqvi, K. Nazarpour, S. Sanei, and J. A. Chambers, "Blind source extraction of heart sound signals from lung sound recordings exploiting periodicity of the heart sound," in *Proc. IEEE ICASSP*, Las Vegas, NV, 2008, pp. 461–464.
[8] 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," in *Proc. EUSIPCO*, Lausanne, Switzerland, 2008.
[9] S. Makino, H. Sawada, R. Mukai, and S. Araki, "Blind separation of convolved mixtures of speech in frequency domain," *IEICE Trans. Fundamentals*, vol. E88-A, no. 7, pp. 1640–1655, 2005.
[10] W. Wang, S. Sanei, and J. A. Chambers, "A joint diagonalization method for convolutive blind separation of nonstationary sources in the frequency domain," in *Proc. ICA*, Nara, Japan, 2003.
[11] S. Ding, J. Huang, D. Wei, and A. Cichocki, "A near real-time approach for convolutive blind source separation," *IEEE Trans. Circuit Syst.–I*, vol. 53, no. 1, pp. 114–128, Jan. 2006.
[12] H. Nguyen and C. Jutten, "Blind source separation for convolutive mixtures," *Signal Process.*, vol. 45, pp. 209–229, 1995.
[13] S. Douglas, *Blind Separation of Acoustic Signals (in Microphone Arrays: Techniques and Applications)*. Berlin, Germany: Springer, 2001.
[14] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, pp. 21–34, 1998.
[15] S. Makino, *Blind Source Separation of Convolutive Mixtures of Speach, (in Adaptive Signal Processing: Applications to Real-World Problems)*. Berlin, Germany: Springer, 2003.
[16] C. Jutten, "Blind separation of sources: An algorithm for sepration of convolitive mixtures," in *Proc. Int. Signal Process. Workshop*, 1992, pp. 275–278, Elsevier.
[17] C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *J. Acoust. Soc. Amer.*, vol. 25, pp. 975–979, Sep. 1953.
[18] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Robust real-time blind source separation for moving speakers in a room," in *Proc. IEEE ICASSP*, Hong Kong, 2003, pp. 469–472.
[19] A. Koutras, E. Dermatas, and G. Kokkinakis, "Blind source separation of moving speakers in real reverberant environment," in *Proc. IEEE ICASSP*, 2000, pp. 1133–1136.
[20] S. M. Naqvi, Y. Zhang, and J. A. Chambers, "A multimodal approach for frequency domain blind source separation for moving sources in a room," in *Proc. IAPR CIP2008*, Santorini, Greece, 2008, pp. 200–204.
[21] R. E. Prieto and P. Jinachitra, "Blind source separation for time-variant mixing systems using piecewise linear approximations," in *Proc. IEEE ICASSP*, 2005, pp. 301–304.

[22] K. E. Hild-II, D. Erdogmus, and J. C. Principe, "Blind source extraction of time-varying, instantaneous mixtures using an on-line algorithm," in *Proc. IEEE ICASSP*, Orlando, FL, 2002, pp. I-993–I-996.
[23] J. Anemuller and T. Gramss, "On-line blind separation of moving sound sources," in *Proc. ICA'99*, 1999.
[24] W. Addison and S. Roberts, "Blind source separtion with non-stationary mixing using wavelets," in *Proc. Int. Conf. Ind. Compon. Anal. Blind Source Separ.*, 2006.
[25] S. Haykin, J. C. Principe, T. J. Sejnowski, and J. McWhirter, *New Directions in Statiatical Signal Processing: From Systems to Brain*. Cambridge, MA: MIT Press, 2007.
[26] B. D. V. Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4–21, Apr. 1988.
[27] P. Aarabi and S. Zaky, "Robust sound localization using multi-source audiovisual information fusion," *Inf. Fusion*, vol. 3, no. 2, pp. 209–223, 2001.
[28] M. Beal, H. Attias, and N. Jojic, "Audio-video sensor fusion with probabilistic graphical models," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2002.
[29] N. Checka, K. Wilson, M. Siracusa, and T. Darrell, "Multiple person and speaker activity tracking with a particle filter," in *Proc. IEEE ICASSP*, 2004, pp. 881–884.
[30] Y. Chen and Y. Rui, "Real-time speaker tracking using particle filter sensor fusion," *Proc. IEEE*, vol. 92, no. 3, pp. 485–494, Mar. 2004.
[31] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," in *Proc. IEEE ICASSP*, 2001, pp. 3021–3024.
[32] J. Fisher, T. Darrell, W. T. Freeman, and P. Viola, "Learning joint statistical models for audio-visual fusion and segregation," in *Proc. Neural Inf. Process. Syst. (NIPS)*, 2000, pp. 772–778.
[33] D. Gatica-Perez, G. Lathoud, I. McCowan, and J. M. Odobez, "A mixed-state i-particle filter for multi-camera speaker tracking," in *Proc. IEEE Int. Conf. Comput. Vision, Workshop Multimedia Technologies for E-Learning and Collaboration (ICCV-WOMTEC)*, 2003.
[34] S. M. Griebel and M. S. Brandstein, "Microphone array source localization using realizable delay vectors," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, 2001.
[35] J. Hershey and J. Movellan, "Real-time speaker tracking using particle filter sensor fusion," in *Proc. Audio Vision: Using Audio-Visual Synchrony to Locate Sounds, in Proc. Neural Inf. Process. Syst. (NIPS)*, 1999.
[36] M. Isard and J. MacCormick, "Bramble: A Bayesian multi-blob tracker," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2001, pp. 34–41.
[37] B. Kapralos, M. Jenkin, and E. Milios, "Audio-visual localization of multiple speakers in a video teleconferencing setting," *Int. J. Imag. Syst. Technol.*, vol. 13, pp. 95–105, 2003.
[38] G. Lathoud and M. Magimai-Doss, "Asector-based, frequency-domain approach to detection and localization of multiple speakers," in *Proc. IEEE ICASSP*, 2005, pp. 265–268.
[39] I. Potamitis, H. Chen, and G. Tremoulis, "Tracking of multiple moving speakers with multiple microphone arrays," *IEEE Trans. Speech Audio Process*, vol. 12, no. 5, pp. 520–529, Sep. 2004.
[40] D. Gatica-Perez, G. Lathoud, J. Odobez, and I. McCowan, "Audio-visual probabilistic tracking of multiple speakers in meetings," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 601–616, Feb. 2007.
[41] D. Sturim, M. Brandstein, and H. Silverman, "Tracking multiple talkers using microphone array measurements," in *Proc. IEEE ICASSP*, 1997, pp. 371–374.
[42] J. Vermaak, M. Gagnet, A. Blake, and P. Perez, "Sequential Monte Carlo fusion of sound and vision for speaker tracking," in *Proc. Int. Conf. Comput. Vis. ICCV*, 2001, pp. 741–746.
[43] B. Vo, S. Singh, and W. K. Ma, "Tracking multiple speakers using random sets," in *Proc. IEEE ICASSP*, 2004, pp. 357–360.
[44] D. B. Ward, E. A. Lehmann, and R. C. Williamson, "Particle filtering algorithms for acoustic source localization," *IEEE Trans. Speech Audio Process*, vol. 11, no. 6, pp. 826–836, Nov. 2003.
[45] F. Talantzis, A. Pnevmatikakis, and A. G. Constantinides, "Audio and visual active speaker tracking in cluttered indoors environments," *IEEE Trans. Syst., Man, Cybern.-B: Cybern.*, vol. 39, no. 1, pp. 7–15, Jan. 2009.
[46] R. Y. Tsai, "A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses," *IEEE J. Robot. Autom.*, vol. RA-3, no. 4, pp. 323–344, 1987.
[47] H. K. Maganti, D. Gatica-Perez, and I. McCowan, "Speech enhancement and recognition in meetings with an audio-visual sensor array," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2257–2269, Nov. 2007.
[48] I. McCowan, M. Hari-Krishna, D. Gatica-Perez, D. Moore, and S. Ba, "Speech acquisition in meetings with an audio visual sensor array," in *Proc. IEEE Int. Conf. Multimedia ICME*, 2005, pp. 1382–1385.
[49] M. Fallon, S. Godsill, and A. Black, "Joint acoustic source localization and orientation estimation using sequential monte carlo," in *Proc. Digital Audio Effects (DAFx-06)*, 2006.
[50] C. Cherry and W. K. Taylor, "Some further experiments upon the recognition of speech, with one and with two ears," *J. Acoust. Soc. Amer.*, vol. 26, no. 4, pp. 554–559, 1954.
[51] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognition (CVPR)*, 2001, pp. 511–518.
[52] B. Ristic, S. Arulampalam, and N. Gordon, *Beyond the Kalman Filter: Particle Filter for Tracking Applications*. Boston, MA: Artech House, 2004.
[53] S. M. Naqvi, Y. Zhang, and J. A. Chambers, "Multimodal blind source separation for moving sources," in *Proc. IEEE ICASSP*, Taipei, Taiwan, 2009, pp. 125–128.
[54] M. Isard and A. Blake, "Condensation: Conditional density propagation for visual tracking," *Int. J. Comput. Vis.*, vol. 29, no. 1, pp. 5–28, 1998.
[55] K. Astom, *Introduction to Stochastic Control Theory*. New York: Academic, 1970.
[56] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ. Press, 2001.
[57] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, *Markov Chain Monte Carlo in Practice*. New York: Chapman & Hall, 1996.
[58] , A. Doucet, N. de Freitas, and N. Gordon, Eds., *Sequential Monte Carlo Methods in Practice*. New York: Springer-Verlag, 2001.
[59] Z. Khan, T. Balch, and F. Dellaert, "An MCMC-based particle filter for tracking multiple interacting targets," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Prague, Czech Republic, 2004, pp. 279–290.
[60] Z. Khan, T. Balch, and F. Dellaert, "MCMC-based particle filtering for tracking a variable number of interacting targets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 11, pp. 1805–1818, Nov. 2005.
[61] J. S. Liu, *Monte Carlo Strategies in Scientific Computing*. New York: Springer-Verlag, 2001.
[62] S. Haykin, *Adaptive Filter Theory*. Upper Saddle River, NJ: Prentice-Hall, 2001.
[63] I. McCowan and H. Bourlard, "Microphone array post-filter based on noise field coherence," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 709–716, Nov. 2003.
[64] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, "Blind source separation based on a fast-convergence algorithm combining ICA and beamforming," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 2, pp. 666–678, Mar. 2006.
[65] L. Parra and C. V. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 7, pp. 352–362, Sep. 2002.
[66] E. Bingham and A. Hyvärinen, "A fast fixed point algorithm for independent component analysis of complex valued signals," *Int. J. Neural Netw.*, vol. 10, no. 1, pp. 1–8, 2000.
[67] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: Wiley, 2001.

Syed Mohsen Raza Naqvi received the First Class B.Eng. degree in industrial electronics engineering from IIEE/NED University of Engineering and Technology, Karachi, Pakistan, in 2001 and the Ph.D. degree in signal processing from Loughborough University, Leicestershire, U.K., in 2009.

Before his postgraduate studies in the U.K., he worked for research and development in Pakistan from January 2002 to September 2005. He is currently working as a Research Associate in the Advanced Signal Processing Group, Electronic and Electrical Engineering, Loughborough University. Prior to that, he worked as a Research Associate in the Centre for Renewable Energy Systems Technology, Electronic and Electrical Engineering, Loughborough University. His research interests include audio-visual speech processing, blind source separation, nonlinear filtering, data fusion, automation and control, embedded system design, and renewable energy.

Miao Yu was born in China in 1986. He received the B.Sc. degree from Shandong University of Science and Technology, Jinan, China, in 2003 and the first-class M.Sc. degree from the Department of Electronic and Electrical Engineering Loughborough University, Loughborough, U.K., in 2008, where, he is currently pursuing the Ph.D. degree with the topic of his research on fall detection for the elderly by exploiting audio and video information.

Mr. Yu won the with the best student award from the Department of Electronic and Electrical Engineering Loughborough University.

Jonathon A. Chambers (S'83–M'90–SM'98) was born in Peterborough, U.K., in 1960. He received the B.Sc. (Hons.) degree in electrical engineering from the Polytechnic of Central London, London, U.K., in 1985, and the Ph.D. degree in digital signal processing from the University of London, London, in 1990.

He was at Peterhouse, Cambridge University, U.K., and the Imperial College London, U.K. From 1979 to 1982, he was as an Artificer Apprentice in Action, Data, and Control in the Royal Navy. He has held academic and industrial positions at Cardiff University, Imperial College London, King's College London, and Schlumberger Cambridge Research, U.K. In July 2007, he joined the Department of Electronic and Electrical Engineering, Loughborough University, Loughborough, U.K., as a Professor of communications and signal processing, and currently leads the Advanced Signal Processing Research Group and a team of researchers in the analysis, design, and evaluation of novel algorithms for digital signal processing with application in acoustics, biomedicine, and wireless communications. He is also the Director of Research in the Department. He has authored or coauthored more than 300 research outputs including two monographs and more than 100 journal articles.

Dr. Chambers is a member of the IEEE Technical Committee on Signal Processing Theory and Methods. He was the Technical Program Chair for the IEEE Workshop on Statistical Signal Processing 2009 and is the Technical Program Co-Chair for ICASSP 2011, Prague. He was the Chairman of the Institute for Electrical Engineers (IEE) Professional Group E5, Signal Processing. He was an Associate Editor for the IEEE SIGNAL PROCESSING LETTERS and for the IEEE TRANSACTIONS ON SIGNAL PROCESSING for two terms. He was awarded the first QinetiQ Visiting Fellowship in 2007 for outstanding contributions to adaptive signal processing.