Video-Aided Model-Based Source Separation in Real Reverberant Rooms

Muhammad Salman Khan, *Student Member, IEEE*, Syed Mohsen Naqvi, *Member, IEEE*,
Ata-ur-Rehman, *Student Member, IEEE*, Wenwu Wang, *Senior Member, IEEE*, and Jonathon Chambers, *Fellow, IEEE*

***Abstract*—Source separation algorithms that utilize only audio data can perform poorly if multiple sources or reverberation are present. In this paper we therefore propose a video-aided model-based source separation algorithm for a two-channel reverberant recording in which the sources are assumed static. By exploiting cues from video, we first localize individual speech sources in the enclosure and then estimate their directions. The interaural spatial cues, the interaural phase difference and the interaural level difference, as well as the mixing vectors are probabilistically modeled. The models make use of the source direction information and are evaluated at discrete time-frequency points. The model parameters are refined with the well-known expectation-maximization (EM) algorithm. The algorithm outputs time-frequency masks that are used to reconstruct the individual sources. Simulation results show that by utilizing the visual modality the proposed algorithm can produce better time-frequency masks thereby giving improved source estimates. We provide experimental results to test the proposed algorithm in different scenarios and provide comparisons with both other audio-only and audio-visual algorithms and achieve improved performance both on synthetic and real data. We also include dereverberation based pre-processing in our algorithm in order to suppress the late reverberant components from the observed stereo mixture and further enhance the overall output of the algorithm. This advantage makes our algorithm a suitable candidate for use in under-determined highly reverberant settings where the performance of other audio-only and audio-visual methods is limited.**

***Index Terms*—Expectation-maximization, reverberation, source separation, spatial cues, time-frequency masking.**

I. Introduction and Related Work

THE objective of source separation systems is to separate individual sources from acoustic mixtures, a task at which humans are adept. The most promising approaches for source separation in the audio-only domain are: frequency-domain convolutive blind source separation (BSS) within which the audio recordings are modeled at each frequency as instantaneous mixtures of the unknown speech sources but the separated signals commonly suffer from the scaling and permutation (arbitrary order of sources) problems; beamforming techniques which tackle the separation problem from a spatial point of view extracting a signal from a specific direction and reducing signals from other directions, but these normally require a greater number of sensors in the array for an improved performance; and methods from the field of computational auditory scene analysis (CASA) in which monaural and binaural cues such as pitch, onset/offset, interaural level difference (ILD), interaural time difference (ITD) are used to enhance source segregation [1]. Time-frequency (TF) masking is a technique used for audio-only source separation motivated both in CASA and BSS which relies on the assumption of signal sparseness i.e., the majority of the TF samples of each signal are almost zero and thus the sources rarely overlap [2]. The TF approach can therefore, unlike conventional BSS algorithms, handle the under-determined problem where the number of sources is greater than the number of sensors, such as in this paper where only two microphones are employed.

Humans, however, perceive sound as a multi-modal process [3], [4] and therefore the authors in [5] proposed a multi-microphone audio-video source separation method showing improvement over audio-only and video-only schemes, but only in an over-determined setting. Sodoyer *et al.* in [6] also proposed to separate an acoustic source by utilizing the coherence with the speaker's lip movements but experimented with mixtures containing only two sources. In [7] they extended this work and utilized audio-video coherence within classical BSS algorithms. They provided separation results for multi-source mixtures but admitted that robustness was limited if the phonetic complexity increased. Their algorithms did not address separation of convolutive mixtures. Rivet *et al.* in [8] presented the combination of audio-visual coherence and BSS to extract speech from convolutive mixtures. Their audio-visual statistical model exploited the relationship between two basic lip visual parameters and the acoustic parameters to solve the indeterminacy problem. They considered only the case of two sources and mixtures. In [9] the authors proposed to separate sources given the facial video of the sources and a synchronous single-microphone recording. They provided separation results for only two sources speaking in front of the camera with limited reverberation. New methods to exploit the visual modality in source separation are therefore required for multi-speaker highly reverberant environments such as teleconferencing or meeting rooms.

Hence, we propose a new source separation algorithm for stereo reverberant mixtures by exploiting cues from video. We model the ILD and the interaural phase difference (IPD) following the approach in [10] and include a model for the mixing

Manuscript received August 30, 2012; revised December 30, 2012, April 15, 2013; accepted April 24, 2013. Date of publication May 06, 2013; date of current version July 12, 2013. This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) of the U.K. under Grants EP/H049665/1 and EP/H050000/1). The work of M. S. Khan was supported by UET, Peshawar, and the Higher Education Commission (HEC) of Pakistan. The work of A.-ur-Rehman was supported by UET, Lahore. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Sharon Gannot.

M. S. Khan, S. M. Naqvi, A.-ur-Rehman, and J. A. Chambers are with the Advanced Signal Processing Group, School of Electronic, Electrical, and Systems Engineering, Loughborough University, Leicestershire, LE11 3TU, U.K. (e-mail: m.s.khan2@lboro.ac.uk; s.m.r.naqvi@lboro.ac.uk; a.ur-rehman@lboro.ac.uk; j.a.chambers@lboro.ac.uk).

W. Wang is with the Centre for Vision, Speech and Signal Processing, Department of Electronic Engineering, University of Surrey, Guildford, GU2 7XH, U.K. (e-mail: w.wang@surrey.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/TASL.2013.2261814

vectors as in [11] where the mean parameter is estimated by utilizing speaker location information obtained from video. The parameters of the probabilistic models are updated iteratively with the EM algorithm. Since the EM algorithm is sensitive to initialization we initialize the direction vector parameter with the location information of the speakers obtained through vision. We compare our method with two audio-only algorithms requiring only two microphones and three other audio-visual algorithms using solely localization cues for separation. We do not make comparison with the methods in [6]–[9] as they require full-frontal close-up views of the speakers' faces, which are not required in our approach. To the best of our knowledge, our algorithm is the first that fuses vision with a CASA-type framework producing time-frequency masks used for speech separation in realistic multi-speaker environments.

In Section II we give an overview of the algorithm and the probabilistic models. Section III explains the video processing and how it is incorporated in the probabilistic speech separation model. In Section IV we discuss the model parameters and the expectation-maximization algorithm. In Section V we provide a variety of simulation results and comparisons confirming the robustness and consistency of our proposed algorithm.

II. Overview of the Proposed Algorithm

An audio-visual source separation algorithm is proposed. Given binaural reverberant mixtures containing at least two sources, the ILD, IPD, and the mixing vectors are modeled probabilistically. The sources are localized in the room using the video process detailed in Section III and the 3-D location estimates are utilized in the probabilistic modeling of the mixing vectors. The optimum model parameters are estimated by the EM algorithm as described in Section IV. Soft time-frequency masks are then formed from the probabilistic modeling to reconstruct each source. The remainder of this section describes the audio processing and the probabilistic models.

Consider a stereo-recorded speech signal with the left and right sensor (ears or microphones) mixture signals denoted as $l(t_s)$ and $r(t_s)$. The mixtures are sampled with the sampling frequency f_a (sampling period $T_a = 1/f_a$) and hence are available at discrete time indices t_s for processing. Assuming that the number of sources I in the mixture is known *a priori*, the convolutive mixing model for both the sensors, as shown in Fig. 1, can be written as $l(t_s) = \sum_{i=1}^{I} s_i(t_s) * h_{li}(t_s)$, and $r(t_s) = \sum_{i=1}^{I} s_i(t_s) * h_{ri}(t_s)$, where $s_i(t_s)$ denote the speech sources, $h_{li}(t_s)$ and $h_{ri}(t_s)$ are the impulse responses associated with the enclosure from source i to the left and right sensor respectively, and $*$ denotes the discrete time convolution operation. The time domain signals are then converted to the TF domain using the short-time Fourier transform (STFT). The interaural spectrogram is obtained by taking the ratio of the STFT of the left and right channels at each time frame t and frequency ω [10] as, $\frac{L(\omega,t)}{R(\omega,t)} = 10^{\alpha(\omega,t)/20} e^{j\phi(\omega,t)}$, where $\alpha(\omega,t)$ is the ILD, measured in dB, and $\phi(\omega,t)$ is the IPD. The IPD observations are constrained to be in the range $[-\pi,\pi)$. We model a source positioned at a certain location with a frequency-dependent interaural time difference (ITD) $\tau(\omega)$, and a frequency-dependent ILD inspired by [10]. The recorded IPD, $\angle\left(\frac{L(\omega,t)}{R(\omega,t)}\right)$ for each TF point, can not always be mapped to the respective τ due to spatial aliasing. The model also requires that τ and the length of $h(t)$ should be smaller than the Fourier transform window. Any portion of $h(t)$ over one window length is considered part of the noise. We adopt the top-down approach described in [10] which makes it possible to map a τ to a recorded IPD at any desired group of frequencies. The phase residual error, the difference between the recorded IPD and the predicted IPD (by a delay of τ samples), in the interval $[-\pi,\pi)$ is defined as, $\hat{\phi}(\omega,t;\tau) = \angle\left(\frac{L(\omega,t)}{R(\omega,t)} e^{-j\omega\tau}\right)$.

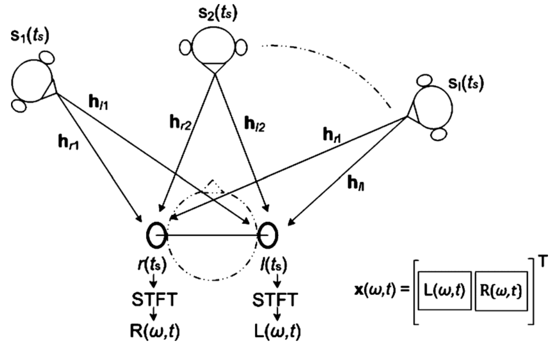


Fig. 1. Signal notations. The left and right sensor convolutive mixtures are transformed to the TF-domain to obtain $\mathrm{L}(\omega,t)$ and $\mathrm{R}(\omega,t)$, and $\mathbf{x}(\omega,t)$ is formed by concatenating $\mathrm{L}(\omega,t)$ and $\mathrm{R}(\omega,t)$ as shown.

The phase residual is modeled with a Gaussian distribution denoted as $p(\cdot)$ with mean $\xi(\omega)$ and variance $\sigma^2(\omega)$ that are dependent on frequency, $p(\phi(\omega,t) \mid \tau(\omega), \sigma^2(\omega)) = \mathcal{N}(\hat{\phi}(\omega,t;\tau) \mid \xi(\omega), \sigma^2(\omega))$. The ILD is also modeled with a Gaussian distribution with mean $\mu(\omega)$ and variance $\eta^2(\omega)$, $p(\alpha(\omega,t) \mid \mu(\omega), \eta^2(\omega)) = \mathcal{N}(\alpha(\omega,t) \mid \mu(\omega), \eta^2(\omega))$. The STFTs of the left and right channels are concatenated to form a new mixture $\mathbf{x}(\omega,t)$ as shown in Fig. 1. Assuming the W-disjoint orthogonality (WDO) property [12] of speech signals, the signals are sparse in the TF domain and only one source is dominant at each TF point, the STFT of the recordings $\mathbf{x}(\omega,t)$ at each time t and frequency ω can be written as [11],

$$\mathbf{x}(\omega,t) = \sum_{i=1}^{I} \mathbf{h}_i(\omega) s_i(\omega,t) \tag{1}$$

and approximated as

$$\mathbf{x}(\omega,t) \approx \mathbf{h}_d(\omega) s_d(\omega,t) \tag{2}$$

where $\mathbf{h}_d(\omega) = [h_{ld}(\omega), h_{rd}(\omega)]^T$ is the mixing vector from the dominant source $s_d(\omega,t)$ to the left and right sensor at that TF point, assumed to be time invariant. To eliminate the effects of source scaling the vector $\mathbf{x}(\omega,t)$ is normalized such that its Euclidean norm is unity and this is performed for each ω and t. The mixing vectors are modeled for each source with a Gaussian model as [11], [13]

$$p(\mathbf{x}(\omega,t) \mid \mathbf{d}_i(\omega), \varsigma_i^2(\omega)) = \frac{1}{\pi\varsigma_i^2(\omega)} \exp\left(-\frac{\left\|\mathbf{x}(\omega,t) - \left(\mathbf{d}_i^H(\omega)\mathbf{x}(\omega,t)\right) . \mathbf{d}_i(\omega)\right\|^2}{\varsigma_i^2(\omega)}\right) \tag{3}$$

where $\mathbf{d}_i(\omega)$ is the direction vector of the direct-path of the source signal which will be obtained from the video measurements, $\varsigma_i^2(\omega)$ is the variance of the model, $(\cdot)^H$ is the Hermitian transpose, and $\|\cdot\|$ indicates the Euclidean norm operator. In [11], [13] and [14] the authors proposed the use of an eigen decomposition of a sample covariance matrix to define unit norm vectors $\mathbf{d}_i(\omega)$ to represent the source directions in the probabilistic modeling of the mixing vectors. This approach, however, will be sensitive to estimation errors due to short data lengths, statistical non-stationarity in the audio scene and background noise. In contrast, in our proposed method the direction vectors are found through vision on the basis of a plane wave assumption, as discussed in Section III-B which thereby overcomes these shortcomings. The resulting TF masks for all sources that are found through the probabilistic modeling will then be improved as explained in Section IV-B. We next explain the video processing and the estimation of the parameter $\mathbf{d}_i(\omega)$.

III. Video Processing

We are proposing an automated audio-visual speech separation algorithm that tracks moving speakers in a room environment and separates their speech mixtures when they are judged to be physically stationary. The moment the speakers enter the monitored environment the video tracker is initialized, and it is then used to automatically follow the speakers to provide the estimates of their locations, and to determine when the sources are essentially static.

For localization of the speakers in a room environment we use at least two fully calibrated color video cameras with overlapping field of view to determine the approximate geometric locations of the speakers. Cameras are calibrated by the Tsai calibration (non-coplanar) technique [15] and synchronized by the external hardware trigger module and frames are captured at the rate of $f_v = 25$ frames/sec [16]. One may argue that audio localization could be used instead of video localization. But in a scenario where multiple speakers are simultaneously active and the environment is highly reverberant the audio localization scheme can fail. Similarly, 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 [17]. Thus, the visual modality with multiple camera integration is chosen as the most suitable approach for speaker localization; combination of audio and video localization is outside the scope of this work. Moreover, the novelty in this paper lies in system integration and the main contribution is to provide video-aided time-frequency masking based speech separation, specifically in under-determined highly reverberant scenarios.

A. Speaker Tracking

The state of each speaker i, where $i = 1, \ldots, I$, in an image at discrete time k is represented as $\mathbf{u}_k^i = [x_k^i, \dot{x}_k^i, y_k^i, \dot{y}_k^i]$, where x_k^i and y_k^i are respectively the x and y coordinates in the image, while $\dot{x}_k^i$ and $\dot{y}_k^i$ are the respective velocities. The combined state of all the I speakers is $\mathbf{U}_k = [\mathbf{u}_k^1, \ldots, \mathbf{u}_k^I]$. Similarly, combined measurements of all the positions of speakers are $\mathbf{Y}_k = [\mathbf{y}_k^1, \ldots, \mathbf{y}_k^I]$.

In Bayesian tracking the main objective is to calculate the posterior probability distribution $p(\mathbf{U}_k \mid \mathbf{Y}_{1:k})$ of the combined state $\mathbf{U}_k$ at discrete time index k. Monte Carlo estimation of the posterior distribution $p(\mathbf{U}_k \mid \mathbf{Y}_{1:k})$ can be represented as

$$p(\mathbf{U}_k \mid \mathbf{Y}_{1:k}) \approx \frac{1}{N_s} p(\mathbf{Y}_k \mid \mathbf{U}_k) \sum_{n=1}^{N_s} p\left(\mathbf{U}_k \mid \mathbf{U}_{k-1}^n\right) \tag{4}$$

where $p(\mathbf{Y}_k \mid \mathbf{U}_k)$ is the likelihood which expresses the measurement model while $p(\mathbf{U}_k \mid \mathbf{U}_{k-1})$ is the prior which expresses the state model and N_s is the number of particles [18].

The Markov chain Monte Carlo based particle filter (MCMC-PF) is used for speaker tracking. The MCMC-PF performs better than the generic PF in multiple speaker tracking scenarios due to the improved sampling efficiency [16], [19]. Multiple speakers can be tracked with a single MCMC-PF while multiple generic particle filters [20] are required to track multiple speakers. The MCMC-PF follows two steps: in the first step we predict a particle to estimate the posterior distribution of the next state and the second step is a refinement step in which we move to accept or reject the predicted particle. The prediction step involves a state model and a suitable proposal distribution, while the refinement step requires a measurement or likelihood model.

Algorithm 1 Video processing for speaker localization

```
Input: Video sequences
Output: 3-D speaker locations
for each camera run the following MCMC-PF do
  Input 2-D positions of the center of the heads and reference patch for each head in the previous frame
  Initialize $N_s$ particles for $I$ number of heads $\{\mathbf{U}_1^n\}_{n=1}^{N_s}$
  for $k = 2, \ldots, K$ do
    Randomly select a particle $r$ from the posterior distribution of the state $\mathbf{U}_{k-1}$ and use this particle and the motion model $q(\cdot)$ to predict the initial state of all the targets at time step $k$
    $\mathbf{U}_k^1 \sim q(\mathbf{U}_k \mid \mathbf{U}_{k-1}^r)$
    for $n = 2, \ldots, N_s + B$ (where $B$ is the number of burn in particles) do
      Randomly select another particle $\mathbf{U}_{k-1}'$ from the posterior distribution at time $k - 1$ $p(\mathbf{U}_{k-1} \mid \mathbf{Y}_{k-1})$
      Propose a new particle using the proposal distribution $Q(\cdot)$ and the randomly selected particle $\mathbf{U}_{k-1}'$
      $\mathbf{U}_k' \sim Q(\mathbf{U}_k^n \mid \mathbf{U}_{k-1}')$
      Compute the measurement likelihoods $p(\mathbf{Y}_k \mid \mathbf{U}_k')$ and $p(\mathbf{Y}_k \mid \mathbf{U}_k^{n-1})$ with respect to the proposed particle $\mathbf{U}_k'$ and the previous particle $\mathbf{U}_k^{n-1}$ respectively
      Compute the acceptance ratio
      $\alpha = min(1, \frac{p(\mathbf{Y}_k \mid \mathbf{U}_k')}{p(\mathbf{Y}_k \mid \mathbf{U}_k^{n-1})})$
      Draw a point $j$ from a uniform distribution
      if $j < \alpha$ then
        $\mathbf{U}_k^n = \mathbf{U}_k'$
      else
        $\mathbf{U}_k^n = \mathbf{U}_k^{n-1}$
      end if
    end for
    Discard the first $B$ particles and keep the remaining of $N_s$ particles.
  end for
end for
Take the estimated 2-D position of the head of speaker $i$ from camera $c$ i.e., $\mathbf{u}^i(c) = [x^i, y^i]$.
Transform $\mathbf{u}^i(1) = [x^i, y^i]$ and $\mathbf{u}^i(2) = [x^i, y^i]$ to 3-D real world Cartesian coordinates $\mathbf{Z}_i$ with the help of camera calibration parameters.
```

1) State Model: To estimate the translation motion of the moving speakers, a constant velocity model [21] is used. The same model is used as a proposal distribution. A rectangular region (patch) which contains the speaker's head is manually selected for each source in the initial frame for which the sources are present in the environment. The pixel in the center of the patch is considered as the location of the mouth. Horizontal and vertical locations of this pixel are tracked in each frame. A 2-D motion of a moving speaker can be described by the constant velocity model [22]

$$\mathbf{u}_{k+1}^{i} = \mathbf{A}\mathbf{u}_{k}^{i} + \mathbf{q}_{k} \tag{5}$$

where $\mathbf{q}_k$ is the measurement noise and the matrix $\mathbf{A}$ is defined as

$$\mathbf{A} = \begin{bmatrix} 1 & T_v & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & T_v \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

where T_v is the frame sampling interval.

In multiple speaker tracking the simple MCMC-PF may fail when a speaker is occluded. Therefore an improved interaction model can also be used to overcome tracking failures and complete details can be found in [23].

2) Likelihood Model: In an MCMC-PF it is very important to have a strong likelihood model. Predicted particles are accepted or rejected on the basis of acceptance ratio α. The likelihood model used is based on the combination of color and gradient histograms [23]. Color histograms are widely used in the literature [22], [24], [25] to exploit the uniqueness of the skin color to track the heads. Scaled versions of red (R), green (G) and blue (B) colors are used in our work. R-G and G-R are used to represent the chrominance information while R+G+B is used to represent the luminance information [26].

Reference histograms H_{ref} are created for all the speakers' heads with the help of the patches selected in the initial frame. For the predicted particles, histograms H_{target} are created by selecting a patch with the predicted state as its center. The Bhattacharyya coefficient ρ between the reference and the target color histograms is calculated by their bin-wise multiplication

$$\rho(H_{ref}, H_{target}) = \sum_{j=1}^{E} \sqrt{H_{ref}^{j} \times H_{target}^{j}} \tag{6}$$

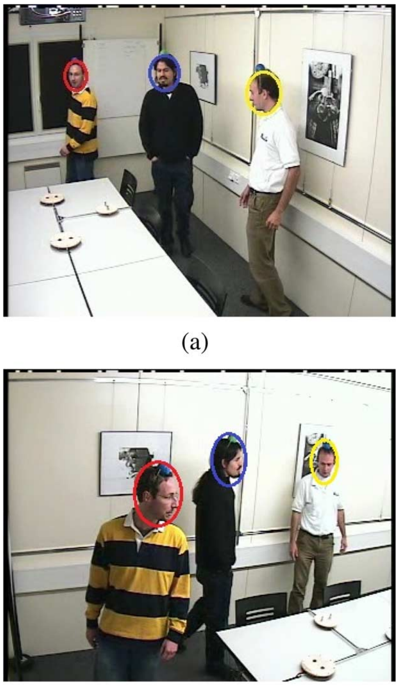
where E represents the number of histograms bins. Bhattacharyya distance [27] between two histograms is defined as

$$d(H_{ref}, H_{target}) = \sqrt{1 - \rho(H_{ref}, H_{target})} \tag{7}$$

The likelihood with respect to the color cues, as in [22] is calculated as

$$L_c\left(\mathbf{y}_k^i \middle| \mathbf{u}_k^i\right) \propto exp\left(-\frac{d\left(H_{ref}^{c}, H_{target}^{c}\right)}{2\sigma^2}\right) \tag{8}$$

where σ^2 is the measurement noise variance.


(a)

(b)

Fig. 2. Images from two cameras with three speakers tracked with an MCMC-PF with an ellipse formed on their heads. The center of the ellipse is assumed to be the approximate 2-D location of the speaker's mouth. The 2-D locations from the images of the two cameras are transformed to 3-D real world Cartesian coordinates. The 3-D coordinates are then used for the $\mathbf{d}_i$ parameter calculation. Note that in this work the cameras are conveniently located at a high level in the room; we remark, however, that similar advantage in video localization could be obtained when the cameras are placed close to each other (and in a similar height to human body), for instance, as in a robotics application, and this topic is an interesting area for future research.

Using only the color histograms is insufficient for tracking purposes because the color based tracker fails when there is something else with a similar color around the speaker. Integration of the gradient histograms [22] helps to overcome such problems. Gradient histograms are created for reference and target patches for the purpose of edge detection. The likelihood with respect to these histograms is calculated by using the Bhattacharyya distance with the help of the following equation

$$L_g\left(\mathbf{y}_k^i \middle| \mathbf{u}_k^i\right) \propto exp\left(-\frac{d\left(H_{ref}^{g}, H_{target}^{g}\right)}{2\sigma^2}\right) \tag{9}$$

where the overall likelihood is then calculated as

$$p\left(\mathbf{y}_k^i \middle| \mathbf{u}_k^i\right) = \nu L_c\left(\mathbf{y}_k^i \middle| \mathbf{u}_k^i\right) + (1-\nu) L_g\left(\mathbf{y}_k^i \middle| \mathbf{u}_k^i\right) \tag{10}$$

and ν is the weighting coefficient, which is used to weight the two video cues.

3) 3-D Speaker Location Estimation: The output of the above MCMC-PF based tracker is the approximate position of the lips of the speaker. The center of the ellipse in image coordinates $\mathbf{u}^i(c) = [x^i, y^i]$, where c represents the camera index, $c = 1, 2$, is assumed to be the approximate 2-D position of the speakers as shown in Fig. 2. The 2-D position information, from images of at least two cameras, are transformed to 3-D real world Cartesian coordinates. In 3-D space each point in each camera frame defines a ray. Intersection of both rays is found by using multi-view geometry, which finally helps in calculation of the location for a speaker $\mathbf{Z}_i = [p_{x_i}, p_{y_i}, p_{z_i}]$ in 3-D real world coordinates [28]. This 3-D geometric information of each speaker is used for $\mathbf{d}_i$ calculation. The video processing for speaker localization is summarized in Algorithm 1.

B. Parameter $\mathbf{d}_i$ *Calculation*

After estimating the 3-D position of each speaker i, the elevation (θ_i) and azimuth (ϕ_i) angles of arrival to the center of the sensors are calculated as $\theta_i = \tan^{-1}\left(\frac{p_{y_i}-p'_{y_c}}{p_{x_i}-p'_{x_c}}\right)$ and $\phi_i = \sin^{-1}\left(\frac{p_{y_i}-p'_{y_c}}{r_i \sin(\theta_i)}\right)$, where $r_i = \sqrt{(p_{x_i}-p'_{x_c})^2+(p_{y_i}-p'_{y_c})^2+(p_{z_i}-p'_{z_c})^2}$, while p'_{x_c}, p'_{y_c} and p'_{z_c} are coordinates of the center of the microphone array. The direct-path weight vector $\mathbf{d}_i(\omega)$ for frequency bin ω and for source of interest (SOI) $i = 1,\ldots,I$, can be derived [29] as:

$$\mathbf{d}_i(\omega) = \begin{bmatrix} \exp(-j\kappa(\sin(\theta_i)\cdot\cos(\phi_i)\cdot p'_{x_1}+\sin(\theta_i)\cdot \\ \sin(\phi_i)\cdot p'_{y_1}+\cos(\theta_i)\cdot p'_{z_1})) \\ \exp(-j\kappa(\sin(\theta_i)\cdot\cos(\phi_i)\cdot p'_{x_2}+\sin(\theta_i)\cdot \\ \sin(\phi_i)\cdot p'_{y_2}+\cos(\theta_i)\cdot p'_{z_2})) \end{bmatrix} \tag{11}$$

where p'_{x_j}, p'_{y_j} and p'_{z_j} for $j = 1,2$ are the 3-D positions of the microphones and $\kappa = \omega/c_s$ and c_s is the speed of sound in air at room temperature. The vector $\mathbf{d}_i(\omega)$ is normalized to unity length before it is used in the model.

To form an accurate time frequency mask for each static source the IPD and ILD models, and the model for the mixing vectors that utilize the direct-path weight vector in (11) obtained with the aid of video are used in conjunction. Since the sources are differently distributed in the mixture spectrograms, in terms of their IPD, ILD and their mixing, the parameters of the above models cannot be obtained directly from those mixtures. It is a hidden maximum-likelihood parameter estimation problem and thus the expectation-maximization algorithm is employed for its solution. Considering the models to be conditionally independent, we combine them given their corresponding parameters as

$$\begin{aligned} p(\alpha(\omega,t),\phi(\omega,t),\mathbf{x}(\omega,t)\,|\,\widetilde{\Theta}) =& \mathcal{N}(\alpha(\omega,t)\,|\,\mu(\omega),\eta^2(\omega)) \\ &\cdot\mathcal{N}(\hat{\phi}(\omega,t)\,|\,\xi(\omega),\sigma^2(\omega)) \\ &\cdot\mathcal{N}(\mathbf{x}(\omega,t)\,|\,\mathbf{d}(\omega),\varsigma^2(\omega)) \end{aligned} \tag{12}$$

where $\widetilde{\Theta}$ denotes all of the model parameters. We emphasize that it is only the noise in the measurements of ILD and IPD that is assumed to be conditionally independent and we adopt this same assumption as in [10] for the measurement related to the source direction vector. However, the conditional independence assumption offers particular advantage in algorithm development; namely, at each iteration of the EM algorithm, the parameters can be updated separately. As in [10], the dependence between ILD and IPD is introduced through prior assumptions on the mean values of the model parameters. Since the ILD and IPD may have dependence on source direction, the assumption of the conditional independence amongst the noise components may only be an approximation. Modeling such dependence is beyond the scope of this study, but is an interesting point for further investigation.

IV. Model Parameters and Expectation-Maximization Algorithm

A. Model Parameters

All of the model parameters $\widetilde{\Theta}$ can be collected as a parameter vector

$$\widetilde{\Theta} = \left\{\mu_i(\omega),\eta_i^2(\omega),\xi_{i\tau}(\omega),\sigma_{i\tau}^2(\omega),\mathbf{d}_i(\omega),\varsigma_i^2(\omega),\psi_{i\tau}\right\} \tag{13}$$

where μ_i, $\xi_{i\tau}$, and $\mathbf{d}_i$ and η_i^2, $\sigma_{i\tau}^2$, and ς_i^2 are respectively the means and variances of the ILD, IPD, and mixing vector models. The subscript i indicates that the parameters belong to the source i, and τ and ω show the dependency on delay and frequency. We include $\mathbf{d}_i(\omega)$ since it is used within the EM algorithm but highlight that since it is obtained from the video it remains constant throughout the algorithm. The parameter $\psi_{i\tau}$ is the mixing weight, i.e., the estimate of the probability of any TF point belonging to source i at a delay τ. Note that $\psi_{i\tau}$ is obtained from the hidden variable $z_{i\tau}(\omega,t)$ [10] that qualifies the assignment of a TF unit to source i for the delay τ. The hidden variable is an important variable and is unity if the TF point belongs to both source i and delay τ and zero otherwise.

The log value of the likelihood function $(\mathcal{L})$ given the observations can be written as

$$\begin{aligned} \mathcal{L}(\widetilde{\Theta}) &= \sum_{\omega,t}\log p(\alpha(\omega,t),\phi(\omega,t),\mathbf{x}(\omega,t)\,|\,\widetilde{\Theta}) \\ &= \sum_{\omega,t}\log\sum_{i,\tau}[\mathcal{N}(\alpha(\omega,t)\,|\,\mu_i(\omega),\eta_i^2(\omega)) \\ &\qquad\cdot\mathcal{N}(\hat{\phi}(\omega,t;\tau)\,|\,\xi_{i\tau}(\omega),\sigma_{i\tau}^2(\omega)) \\ &\qquad\cdot\mathcal{N}(\mathbf{x}(\omega,t)\,|\,\mathbf{d}_i(\omega),\varsigma_i^2(\omega))\cdot\psi_{i\tau}] \end{aligned} \tag{14}$$

and the maximum likelihood solution is the parameter vector which maximizes this quantity.

B. The Expectation-Maximization Algorithm

The algorithm is initialized with the estimated directions of the speakers provided by video and the PHAT histogram. In the expectation step (E-step) the probabilities are calculated given the observations and the estimates of the parameters as

$$\begin{aligned} \epsilon_{i\tau}(\omega,t) =& \psi_{i\tau}\cdot\mathcal{N}\left(\alpha(\omega,t)\,|\,\mu_i(\omega),\eta_i^2(\omega)\right) \\ &\cdot\mathcal{N}\left(\hat{\phi}(\omega,t;\tau)\,|\,\xi_{i\tau}(\omega),\sigma_{i\tau}^2(\omega)\right) \\ &\cdot\mathcal{N}\left(\mathbf{x}(\omega,t)\,|\,\mathbf{d}_i(\omega),\varsigma_i^2(\omega)\right) \end{aligned} \tag{15}$$

where $\epsilon_{i\tau}(\omega,t)$ is the expectation of the hidden variable. In the maximization step (M-step), the parameters are updated using the observations and $\epsilon_{i\tau}(\omega,t)$ from the E-step. The IPD and ILD parameters and $\psi_{i\tau}$ are re-estimated as in [10]. The mean parameter of the mixing vectors $\mathbf{d}_i(\omega)$ is obtained through video as discussed in Section III-B and $\varsigma_i^2(\omega)$ is updated as [11]

$$\varsigma_i^2(\omega) = \frac{\sum_{t,\tau}\epsilon_{i\tau}(\omega,t)\cdot\left\|\mathbf{x}(\omega,t)-\left(\mathbf{d}_i^H(\omega)\,\mathbf{x}(\omega,t)\right)\cdot\mathbf{d}_i(\omega)\right\|^2}{\sum_{t,\tau}\epsilon_{i\tau}(\omega,t)}. \tag{16}$$

The mixing vector model starts contributing from the second iteration, as in the first iteration the occupation likelihood $\epsilon_{i\tau}(\omega, t)$ is calculated using only the ILD and IPD models. The initial value of $\varsigma_i^2(\omega)$ is computed after the first iteration using $\epsilon_{i\tau}(\omega, t)$. Since the algorithm is initialized with source locations estimates from video and $\epsilon_{i\tau}(\omega, t)$ contains the correct order of the sources the permutation problem is bypassed. The probabilistic masks for each source can be formed as $M_i(\omega, t) \equiv \sum_\tau \epsilon_{i\tau}(\omega, t)$. The time domain source estimates are obtained by applying the TF masks to the mixtures and taking the inverse STFT. We next experimentally verify the efficacy of the proposed approach. A brief summary of the proposed scheme is given in Algorithm 2.

Algorithm 2 Brief summary of the proposed audio-visual source separation approach

```
Input: Synchronized audio-visual measurements
Output: Separated speech sources
1: Run Algorithm 1 to obtain the speaker locations when
   the sources are judged physically stationary
2: Calculate parameter $\mathbf{d}_i$ as in Section III-B
3: Initialize certain parameters of the EM algorithm in
   Section IV-B with speaker locations and the PHAT
   histogram
4: Run the EM algorithm as in Section IV-B to generate
   time-frequency masks for all sources
5: Apply the time-frequency masks to the mixtures to
   reconstruct the sources
```

V. Experimental Evaluation in a Room Environment

We evaluate the performance of the proposed algorithm in two main sets of experiments and compare it with five other algorithms, two are audio-only and three are audio-visual. Firstly, we simulate mixtures of two sources with varying reverberation times (RT60s) using synthetic room impulse responses (RIRs), different model complexities and separation angles, and three sources with varying separation angles utilizing real RIRs. We also conduct experiments on the AV16.3 audio-visual corpus [30] containing real room recordings. We provide comparisons in all of the above scenarios with two other state-of-the-art audio-only algorithms to highlight the advantage of the audio-visual approach to source separation. Secondly, we perform experiments for varying RT60s for both two and three source mixtures and compare the proposed method with three other state-of-the-art audio-visual algorithms.

A. Common Experimental Settings

1) Room Layout: The room setting is shown in Fig. 3. Experiments were performed for mixtures of both two and three speech sources. The desired source was located in front of the sensors at 0° azimuth and the interferer was positioned at one of the six different azimuths between 15° and 90° i.e., [15°, 30°, 45°, 60°, 75°, 90°] for the case of two speakers. In the three-speaker case the third source was located symmetrically with the same azimuth, as shown for approximately 60° in Fig. 3.

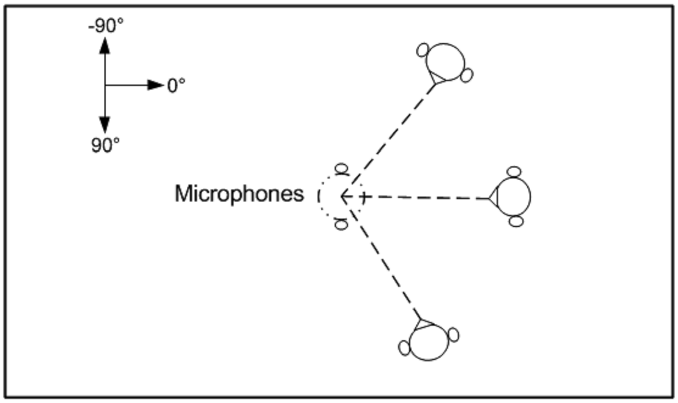


Fig. 3. The room layout showing one of the approximate positions of the sources and the sensors.

2) Parameters Used in Video Processing: When processing the video sequences from the AV16.3 corpus the total number of particles used in the MCMC-PF is $N_s = 400$ with a burn in period of $B = 100$. From the experimental results it is observed that in most of the cases the color cues perform better than the gradient cues, so more weight is given to the color cues by setting $\nu = 0.7$ and $16 \times 16 \times 16$ histogram bins are used for likelihood modeling. The measurement noise variance $\sigma^2 = 0.001$. These parameters were chosen empirically to optimize performance.

3) Speech Data and Room Impulse Responses: Speech signals from the TIMIT acoustic-phonetic continuous speech corpus [31] were used. We randomly chose utterances to form mixtures with different combinations i.e., male-male, male-female, and female-female. The first (16 k×2.5) samples of the TIMIT speech sources were used and were normalized to unity variance before convolving with the RIRs. Both real and simulated RIRs were used. The real RIRs come from [32] which were measured in a real classroom with an RT60 of approximately 565 ms. We used the center location in our experiments and the sensor-to-speaker distance of 1 m. The simulated RIRs were generated by the image method [33] to evaluate our algorithm for varying RT60s.

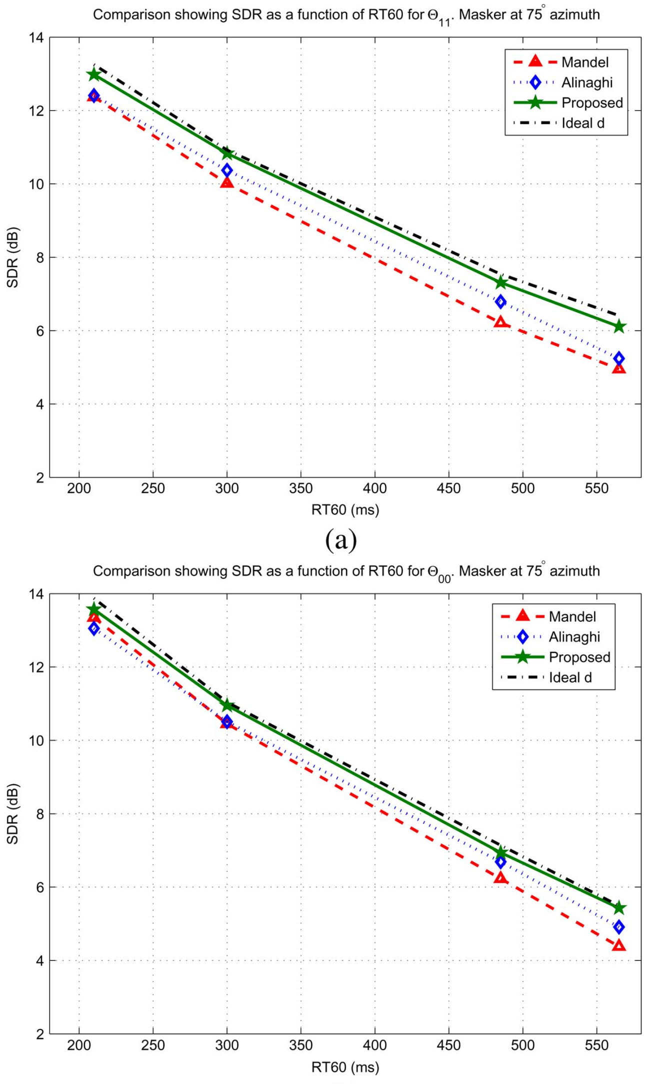
4) Evaluation of Separation Performance: The signal-to-distortion ratio (SDR) as in [34] was used to evaluate the performance of our algorithm in cases where the original speech sources were available. For the AV16.3 corpus since only the real microphone (mixture) measurements are available and we do not have access to the original speech sources, the performance can not be evaluated using [34]. We thus use pitch as a feature to compare separation performance [35], considering the fact that speech sections at different time slots have different pitches [36] and given that the original sources do not have substantially overlapping pitch characteristics. The pitch difference is given as, $p_{diff}(t) = \sqrt{\sum_{i \neq j}(p_i(t) - p_j(t))^2}, \quad i, j = 1, \cdots, m$, and $t = 1, \cdots, T_p$, where T_p is the number of time slots. If the pitch difference is greater than a threshold p_{thr} at a certain time slot, the mixed signals are considered separated at that time slot and the separation status $sep\_status(t)$ is set to unity, otherwise zero. The separation rate is then calculated to evaluate the separation performance as $separation\_rate = \frac{\sum_t sep\_status(t)}{T_p}$. The separation performance improves as the separation rate increases. It is highlighted that objective evaluations for real mixtures can not portray the true quality of the separated speech signals, although they can be used to compare the separation performance of different separation methods. We therefore also conduct listening tests and provide mean opinion scores (MOS tests for voice are specified by ITU-T recommendation P.800 which are followed in our evaluation) for the AV16.3 dataset.

B. Results and Comparison With Other Audio-Only Algorithms

Extensive experiments were conducted to test the robustness and consistency of our proposed algorithm. The common parameters used in all experiments are given in Table I. As mentioned earlier, to emphasize the advantage of our multimodal approach over audio-only methods in realistic multi-speaker environments we compare our results with [10], which we refer to as Mandel, and [14], which we term Alinaghi. Note that in the simulations where speech from TIMIT is convolved with RIRs to generate the speech mixtures, we have avoided using the information from the video system, in order to perform comparison with other methods. In the "ideal" case, referred to in the results as "Ideal d", the exact locations of the sources and the microphones, which were used in the generation of the RIRs, are used to calculate the exact direction of arrival (DOA) of each source to the center of the microphone array. From the localization results by the video tracking algorithms [16], [37], as described in Section III-A, we have observed that the DOA information estimated from the video recordings would contain estimation errors. Therefore, in the case of simulated room environments where the mixtures are generated using the sources from TIMIT, video information was not used. Instead, in the "Proposed" method we use the exact DOAs of the sources perturbed by zero-mean Gaussian noise with a standard deviation of 3 degrees, which corresponds approximately to the average of that for the three speakers given in Fig. 5 of [37]. Such a simulation set-up applies to the results in Figs. 4–6, 8, and 10.

Different model complexities, for ILD and IPD, were evaluated similar to [10]. For instance, the ILD and IPD model complexity of Θ_{00} will have no ILD contribution and an IPD model with zero mean and a standard deviation that varies only by source, whereas Θ_{11} will have a frequency-independent ILD model and an IPD model with a frequency-independent mean and a standard deviation that varies by source and τ, while $\Theta_{\Omega\Omega}$ uses the full frequency-dependent ILD and IPD model parameters. And $\Theta_{\Omega\Omega}^{G}$ has parameters similar to $\Theta_{\Omega\Omega}$ but includes a garbage source and an ILD prior as described in [10].

In Fig. 4, the two model complexities Θ_{11} and Θ_{00} for two sources were simulated with an interferer at 75°. The speech files from the TIMIT dataset were convolved with the RIRs generated using the image method [33] to obtain the reverberant mixtures. The RT60 was varied to evaluate performance of the algorithms at different levels of reverberation. A curve that corresponds to the model which uses the ideal $\mathbf{d}_i$ vector found from the known source locations has also been included in the results. The curve provides an upper bound for performance improvement for the algorithm. The results indicate the improved performance of our proposed technique over [10] and [14]. In Fig. 4(a), for RT60 of 210 ms our algorithm gives an output of 12.98 dB, Mandel's algorithm gives 12.37 dB and Alinaghi 12.41 dB. As the RT60 increases our algorithm still performs best, for example at 565 ms it is 6.11 dB, which is 1.16 dB higher than Mandel and 0.87 dB higher than the method by Alinaghi. In Fig. 4(b), with a simpler model Θ_{00}, at an RT60 of 210 ms our method outputs 13.57 dB, compared to Mandel, 13.35 dB, and Alinaghi, 13.05 dB. At the maximum RT60 of 565 ms our algorithm gives an output of 5.43 dB, 1.05 dB higher than Mandel and 0.52 dB higher than Alinaghi. The ILD cues fade away with increasing reverberation and thus the direct-path direction vector obtained by video information in the proposed algorithm contributes to better model the mixing vectors and improve the separation performance.

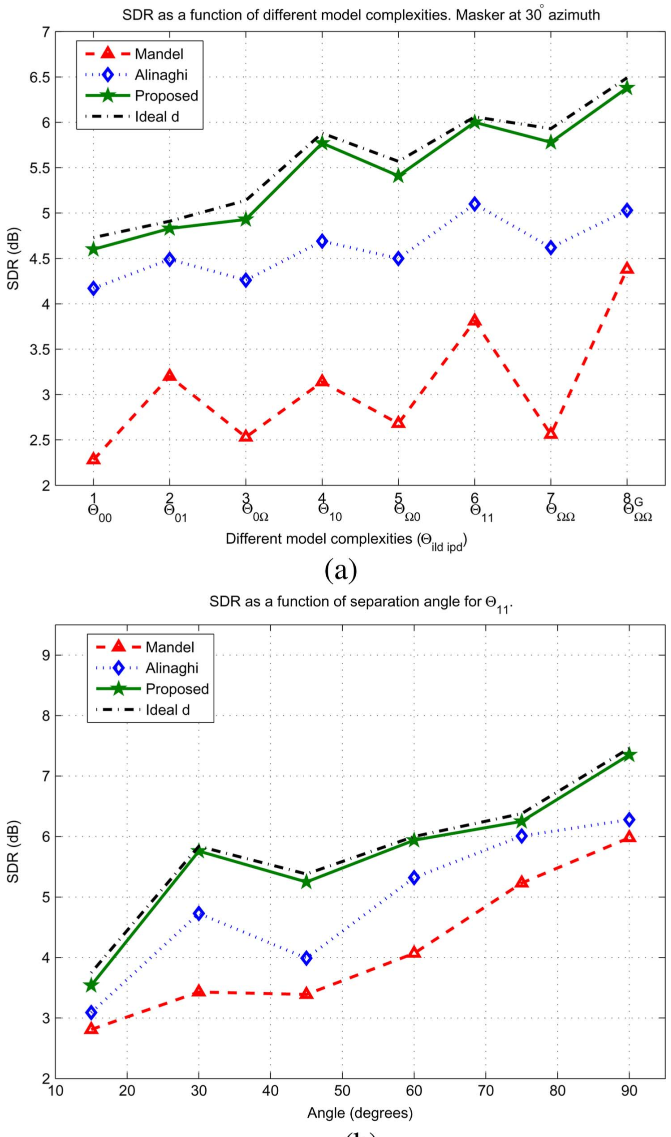


(a)

(b)

Fig. 4. Comparison of performance at different RT60s. The interferer was located at 75° azimuth. Synthetic RIRs using [33] were used to simulate varying RT60s. The Θ_{11} (a) and Θ_{00} (b) modes are under consideration.

In Fig. 5(a) our proposed algorithm was evaluated for all the model complexities. Real RIRs from [32] were utilized to form acoustic mixtures in this set of experiments. The results indicate that our algorithm's performance is consistently better than the compared methods for all models. In [14] the authors reported that their algorithm showed significant improvement over [10] with simpler models but the improvement diminished with the increasing model complexity as confirmed in Fig. 5(a),



(a)

(b)

Fig. 5. In (a) the performance at different model complexities $\Theta_{ild\ ipd}$ for two sources with the interferer at 30° azimuth is shown. The graph in (b) indicates results at different separation angles for model Θ_{11}. The position of the interferer was varied in steps of 15° between 15° to 90°. Real binaural RIRs from [32] were used. Results were averaged over five random mixtures. Our proposed method yields a considerable improvement at all modes and separation angles.

specifically when the ILD model started contributing. In contrast, the performance of our algorithm is clearly shown not to deteriorate with increasing complexity and shows consistent improvement over all the models. The average improvement across the models in the Alinaghi method over the Mandel method is 1.53 dB, whereas for our method is 2.39 dB. In Fig. 5(b) the SDR as a function of the separation angle between the speakers for the Θ_{11} model is shown. Comparatively, over all angles our algorithm that utilizes the estimate of the source direct-path direction vector, by exploiting visual information, yields an average improvement of 1.53 dB whereas Alinaghi's method gives 0.75 dB over Mandel's method.

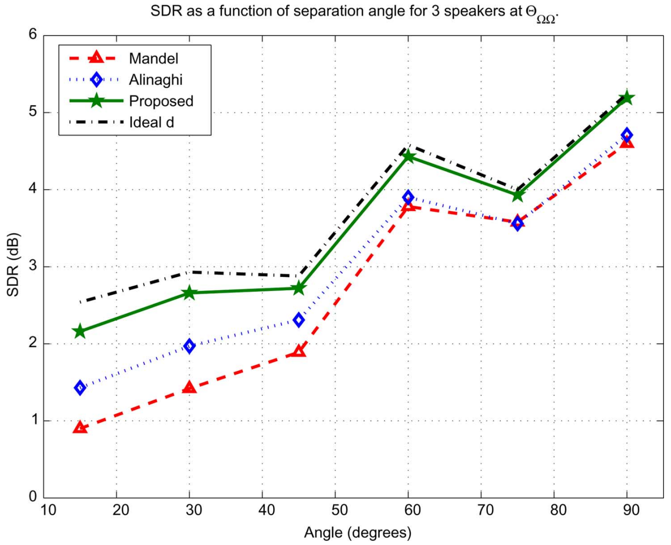


Fig. 6. Results of the three-speaker case at different separation angles using the real RIRs at the $\Theta_{\Omega\Omega}$ mode. The interferers were located symmetrically to both sides of the target source. Results indicate that our proposed method performs best at all separation angles.

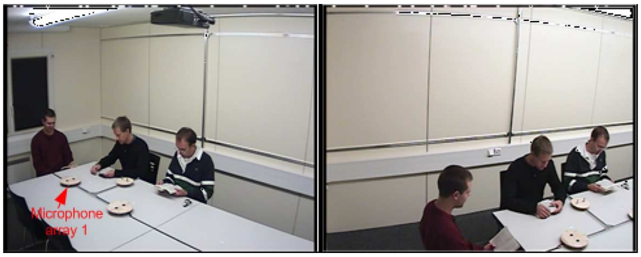


Fig. 7. Image from camera 1 on the right and camera 2 on the left. All the three speakers are seated and simultaneously active for the time slots under consideration. Mixtures from the third and seventh sensor of the microphone array 1 were utilized.

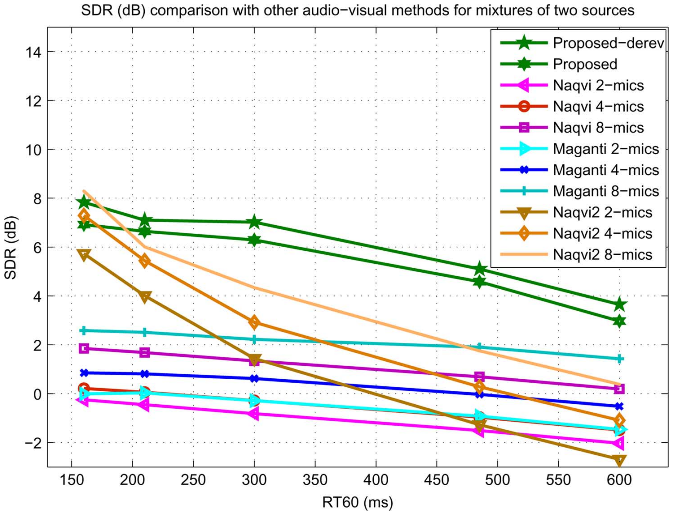


Fig. 8. Comparison of SDR (in decibels) performance as a function of RT60 using the proposed algorithm with and without dereverberation utilizing two microphones and the Naqvi, Maganti and Naqvi2 methods employing two, four and eight microphones for mixtures of two sources.

Results in Fig. 6 show SDR as a function of separation angle i.e., between 15° and 90° for mixtures of three speakers with the most complex frequency-dependent mode $\Theta_{\Omega\Omega}$ using real RIRs. The two interferers on either side of the target were positioned symmetrically with the same azimuth. The interferer to the left was simulated by reversing the order of the sensors. At the minimum separation angle of 15° our algorithm gives an output

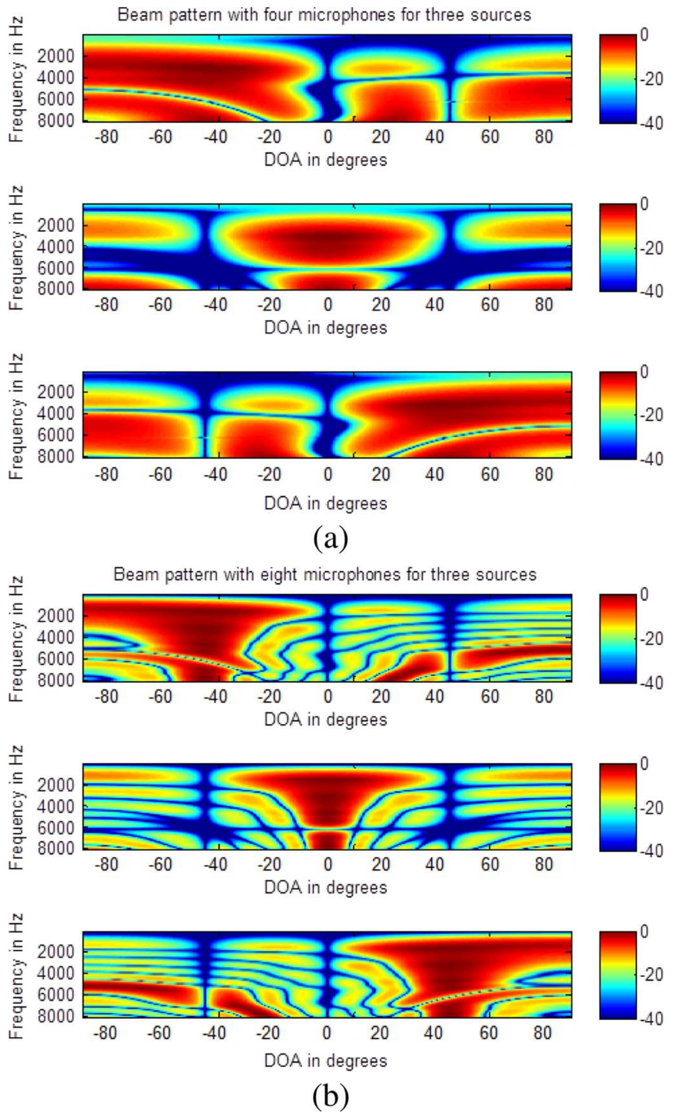


Fig. 9. Beam patterns achieved by the beamformer in Naqvi2 with four microphones in (a) and eight microphones in (b) for the case of three sources. It is clearly visible that as the number of sensors is increased the beam for the desired source becomes more precise strictly allowing the desired source and forming a null towards the interferer. With fewer microphones the interferers and reverberation leak through with the desired source degrading the separation performance.

of 2.16 dB, whereas Mandel, 0.9 dB, and Alinaghi, 1.43 dB. The results indicate that the method in [14] offers improvement over [10] at smaller separation angles from 15° to 45° but no significant improvement at larger separation angles. Our proposed algorithm, in contrast, shows consistent improvement over all separation angles, specifically in the difficult scenario with smaller separation angles, over both [10] and [14] in the three-speaker reverberant case confirming the suitability of our audio-visual approach in multi-speaker realistic settings, and the value of adding visual information in audio source separation.

1) Results for the AV16.3 Corpus: The AV16.3 audio-visual corpus [30] has indoor multi-speaker recordings of real human speakers. We used data from the case where three speakers were seated and were simultaneously active. The direction vector $\mathbf{d}_i(\omega)$ was obtained as explained in Section III-B, which relies upon using the speaker tracking described in Section III-A, and was used in the model. Speech mixtures from the third and seventh sensor of the microphone array 1 as in Fig. 7 were utilized. The mixtures have a duration of eight minutes but we used only the time slots when all the three speakers were active and we selected three seconds data from it. Table II summarizes the pitch-based separation rates and MOSs (six people participated in the listening tests) for the proposed algorithm compared with [10] and [14]. In results, for instance, over the time slot 214–217 s the proposed algorithm has a separation rate of 0.223 and an MOS of 4.2, higher than [10], 0.216 and 3.9, and [14], 0.220 and 4.0, confirming that our proposed algorithm improves the separation performance on real room recordings.

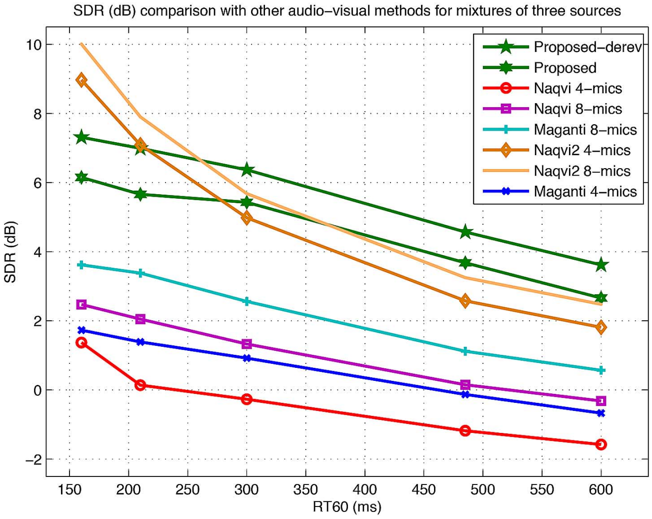


Fig. 10. Comparison of SDR (in decibels) performance as a function of RT60 using the proposed algorithm with and without dereverberation utilizing two microphones and the Naqvi, Maganti and Naqvi2 methods employing four and eight microphones for mixtures of three sources.

TABLE I
DIFFERENT PARAMETERS USED IN EXPERIMENTS

STFT frame length	1024
Velocity of sound	343 m/s
Reverberation time	565 ms (real) or 160-600 ms (image method)
Room dimensions	[9 5 3.5] m or [8.2 3.6 2.4] m (AV16.3)
Source signal duration	2.5 s (TIMIT) or 3 s (AV16.3)
Sensor spacing	0.17 m or 0.2 m (AV16.3)

TABLE II
SEPARATION RATES AND MOSs FOR DIFFERENT TIME SLOTS OF THE AV16.3 CORPUS FOR THE THREE-SPEAKER CASE

Time slot	Mandel		Alinaghi		Proposed	
(seconds)	Separation rate	MOS	Separation rate	MOS	Separation rate	MOS
214-217	0.216	3.9	0.220	4.0	0.223	4.2
218-221	0.093	3.7	0.080	3.7	0.096	3.9
230-233	0.260	3.9	0.253	3.9	0.266	4.1
247-250	0.340	3.9	0.353	4.0	0.363	4.3

C. Comparison With Other Audio-Visual Methods

We next compare our approach with three other audio-visual algorithms, the beamforming based method in [16] which we refer to as Naqvi, the technique in [17], which we term as Maganti and the scheme in [38] using robust beamforming, which we refer to as Naqvi2. Similar to our work, these audio-visual methods employ the visual modality to estimate the speaker locations which are then utilized within the algorithms. For evaluation on the sequences from the AV16.3 corpus, the speech mixtures from the third and seventh sensor of the microphone array 1 as in Fig. 7 were utilized for evaluation of our approach and mixtures from all eight sensors of the same array were used for the above mentioned three beamforming based audio-visual techniques.

The multimodal approach to blind source separation [16] uses the visual modality to enhance the separation of both static and moving sources. The speaker positions estimated by a 3-D tracker are used to initialize the frequency domain BSS algorithm for the physically stationary speakers and beamforming if the speakers are moving. The algorithm's performance is reasonable at low reverberation when the direct path signal is strong but deteriorates at higher RT60s when the direct-to-reverberant ratio (DRR) is low. The beamformer is also generally limited to the determined and over-determined cases and achieves improved performance with larger number of audio sensors.

In [17] an audio-video multi-speaker tracker is proposed to localize sources and then separate them using microphone array beamforming. A postfiltering stage is then applied after the beamforming to further enhance the separation. The overall objective of the system is automatic speech recognition which lies outside the scope of our work, thus, we compare the output of the speech enhancement part, which we implemented ourselves.

In [38] a robust least squares frequency invariant data independent beamformer is implemented. The MCMC-PF based tracker estimates the direction of arrival of the sources using visual images obtained from at least two cameras. The robust beamformer, given the spatial knowledge of the speakers, uses a convex optimization approach to provide a precise beam for the desired source. To control the sensitivity of the beamformer a white noise constraint is used. The scheme provides significant improvement at lower RT60s but the performance degrades as reverberation increases. We employ the original code used in [38] in our comparison.

In contrast, in [6] a speech source is separated by utilizing its coherence with the speaker's lip movements. Parameters describing a speaker's lip shape are extracted using a face processing system. The authors provide results for separation of simple vowel-plosive combinations from other meaningful utterances and acknowledge that separating complex mixtures would be increasingly difficult. In the extension of their work in [7], the spectral content of the sound that is linked with coherent lip movements is exploited and assessment is provided on two audio-visual corpora, one having vowel-plosive utterances similar to their previous work and the other containing meaningful speech spoken by a French speaker. They discuss the determined case and the under-determined case with two sensors and three sources but reported that performance was limited as the phonetic complexity increased. These works, as in [8], [9], require the speakers to be right in front of the camera(s), with the face clearly visible so that facial cues can be observed. Our approach is more general, in that only head localization information is required and therefore audio-visual recordings with low resolution (such as the AV16.3) can be processed, as in Figs. 2 and 7. Hence we do not include the methods in [6]–[9] in our comparison.

1) Pre-Processing for Dereverberation: To reduce the effects of reverberation from the observed mixture, we also employ a pre-processing stage based on spectral subtraction before applying our proposed algorithm and include its results in the comparisons. This dereverberation scheme is based on the RIR modeling proposed for the single-channel case in [39], which was extended to the binaural context in [40]. We implement the first stage of [40] ourselves that suppresses the effects of the late reflections. This pre-processing suits our algorithm well, since it explicitly preserves the binaural cues, ILD and IPD, that are exploited in our proposed separation algorithm. The spectral gain smoothing as in [40] is applied since musical noise is introduced in the processing. All the parameter values were chosen as proposed by the authors of the original paper. We do not estimate the reverberation time in this work and assume it is known.

2) Results: The experimental results in Fig. 8 provide the average SDR (dB) as a function of RT60 for ten random mixtures of two sources for the proposed method and the three other audio-visual methods i.e., Naqvi, Maganti, and Naqvi2. The masker was positioned at -15 degrees azimuth i.e., the minimum and most challenging separation angle in the earlier simulations. The other algorithms were each evaluated with two, four and eight microphones at all RT60s. The proposed algorithm with and without the pre-processing gives better separation, using only two microphones, than all the other algorithms at all RT60s except at 160 ms where the Naqvi2 outperforms the proposed method with four microphones and the proposed algorithm with the pre-processing, Proposed-derev, with eight microphones. The Naqvi and Maganti methods adopt the general trend by improving the separation as the number of microphones is increased, since the increased number of filter coefficients provides better interference removal. The postfiltering stage in Maganti's scheme refines the output further from its previous beamforming stage by exploiting sparsity of the speech sources. Masking postfilters are obtained by retaining the maximum filter output values at each frequency bin. The final postfilter is then applied to the beamformer output. This scheme considerably improves the performance over that of Naqvi for all number of microphones and all RT60s in terms of the SDR, but introduces musical noise which we observed when we listened to the reconstructed source. In the Naqvi2 method the designed unmixing filters we use are frequency invariant and data independent thus the source statistics and RT60 are not considered. Also, since the physical separation between the sources is only 15°, the increased spatial selectivity of the Naqvi2 design appears to deteriorate the separation performance at higher RT60s. In summary, among the other three competing techniques, "Naqvi2 8-mics" has the best performance for RT60 below 450 ms and "Maganti 8-mics" performs best for RT60 over 450 ms.

The results in Fig. 10 show the average SDR (dB) as a function of RT60 for ten random mixtures for the proposed method and the three other audio-visual methods when separating three sources. Each of these three algorithms was run by using four and eight microphones. Having three sources in the mixture, the case of only two-microphones becomes under-determined and solution is not possible through the beamformers in Naqvi, Maganti, and Naqvi2, unlike the proposed algorithm which can handle the under-determined case too. The separation performance of the Proposed-derev is clearly better than the other algorithms at all the RT60s except for the Naqvi2 at RT60 of below around 210 and 250 ms with four and eight microphones respectively. The improved spatial selectivity of the Naqvi2 design again explains this advantage but this degrades with increasing RT60. All the algorithms follow this general trend of degraded performance with increased RT60. For 160 ms and 210 ms utilizing the eight microphones mixture Naqvi2 performs best. This is the strength of the Naqvi2 method that at lower RT60s, with reduced reflections, and hence fewer reflections from the interfering source and overall reverberation leak through the precise beam formed for the desired source, the separation performance is greatly enhanced. This behavior changes as the RT60 increases beyond 300 ms, where even increasing the number of microphones does not stop the deterioration in the separation performance of the beamformer. In Fig. 9, as an example, the beam pattern for the Naqvi2 beamformer is provided using four and eight microphones for the case of three sources. The sources are positioned at -45°, 0°, and 45°. The beam towards the desired source becomes more precise as the number of microphones is increased. Note, that for Fig. 8 the masker is at -15° which explains why separating three sources can be better with beamforming.

3) Results for the AV16.3 Corpus: The results of the experiments on the AV16.3 dataset can be seen in Table III where similar to Section V-B.1 all the audio-visual algorithms are evaluated according to their pitch-based separation rates (S.R.) and MOSs. In terms of the S.R., the proposed method without pre-processing performs better than all the other algorithms over all the considered time slots. The Proposed-derev algorithm performed second best for the first and last time slot, while the Naqvi approach is the second best for the second and third slots. The Naqvi2 was consistently ranked fourth, and although Maganti did a fair job in isolating the sources it performed worse in this pitch-based objective measure. We believe that the spectral postfiltering within Maganti's scheme disturbs the pitch information, giving a zero in the time slots from 214–217 s and 230–233 s, for the pitch based evaluation metric. In the Proposed-derev algorithm, we believe, the spectral gain smoothing [40] applied to reduce the musical noise produced due to the spectral subtraction at the pre-processing, holds the pitch cues intact helping it to perform second best for the first and last time slots and third for ranges 218–221 s and 230–233 s. The weakness of the pitch based evaluation metric is that it decides the separation performance solely on the pitch content of the separated speech, which if distorted can not give a full picture of how well the sources separated. Unlike the pitch based S.R.'s, the MOS results did not fluctuate much from the algorithms' behavior on synthetic data, making it a more consistent measure compared with the pitch based evaluation. The separation by the Proposed-derev algorithm clearly did better in reducing the reverberation and smoothing the musical noise. Separation by the Maganti was strongly affected by musical noise, although the sources were fairly isolated. Separations by Naqvi and Naqvi2 followed each other closely and were consistent exploiting mixtures from eight microphones.

TABLE III
COMPARISON OF SEPARATION RATES AND MOSs FOR DIFFERENT TIME SLOTS OF THE AV16.3 CORPUS FOR THE THREE-SPEAKER CASE

Time slot	Proposed-derev		Proposed		Naqvi		Maganti		Naqvi2	
(seconds)	S.R.	MOS	S.R.	MOS	S.R.	MOS	S.R.	MOS	S.R.	MOS
214-217	0.096	4.1	0.223	4.1	0.064	3.8	0	3.3	0.046	4.0
218-221	0.050	3.8	0.096	3.9	0.073	3.6	0.029	3.5	0.039	3.8
230-233	0.096	3.9	0.266	4.0	0.159	3.4	0	3.1	0.043	3.8
247-250	0.263	4.2	0.363	4.3	0.152	3.9	0.006	3.5	0.116	3.9

VI. CONCLUSION

By utilizing information from video, we have confirmed that more accurate TF masks can be obtained which give improved source estimates, particularly in highly reverberant multi-speaker environments. We have experimentally tested our proposed system in a variety of settings including for the first time real audio-visual data confirming its robustness over two other audio-only methods and three similar audio-visual algorithms in both the two-speaker and three-speaker cases. We emphasize that this study was to demonstrate the advantage of exploiting information from video in CASA-type time-frequency audio source separation with a fixed number of microphones; complexity issues and real-time implementation fall outside of the scope of this study.

Future work will focus on improving the dereverberation based pre-processing used for the suppression of late reverberant components in the observed mixture, fusing the audio-visual modalities in tracking moving speakers, and processing a changing number of sources within the monitored environment.

ACKNOWLEDGMENT

The authors would like to thank M. Mandel for sharing his Matlab code, Y. Liang for his help in the pitch based evaluation, T. Jan for his help in the implementation of the dereverberation algorithm, and B. Shinn-Cunningham for providing the real BRIRs. We also wish to thank the Associate Editor and the anonymous reviewers for their insightful comments which have considerably improved the clarity of our paper.

REFERENCES

[1] N. Roman, D. Wang, and G. J. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Amer.*, no. 114, pp. 2236–2252, 2003.

[2] D. S. Brungart, P. S. Chang, B. D. Simpson, and D. Wang, "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," *J. Acoust. Soc. Amer.*, vol. 120, pp. 4007–4018, 2006.

[3] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *J. Acoust. Soc. Amer.*, vol. 26, no. 2, pp. 212–215, 1954.

[4] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.

[5] K. Wilson, N. Checka, D. Demirdjian, and T. Darrell, "Audio-video array source separation for perceptual user interfaces," in *Proc. Workshop Perceptive User Interfaces*, Orlando, FL, USA, 2001, pp. 1–7.

[6] D. Sodoyer, J. Schwartz, L. Girin, J. Klinkisch, and C. Jutten, "Separation of audio-visual speech sources: A new approach exploiting the audio-visual coherence of speech stimuli," *EURASIP J. Appl. Signal Process.*, vol. 2002, no. 1, pp. 1165–1173, 2002.

[7] D. Sodoyer, L. Girin, C. Jutten, and J. Schwartz, "Developing an audio-visual speech source separation algorithm," *Speech Commun.*, vol. 44, no. 1–4, pp. 113–125, 2004.

[8] B. Rivet, L. Girin, and C. Jutten, "Mixing audiovisual speech processing and blind source separation for the extraction of speech signals from convolutive mixtures," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 96–108, Jan. 2007.

[9] A. L. Casanovas, G. Monaci, P. Vandergheynst, and R. Gribonval, "Blind audiovisual source separation based on sparse redundant representations," *IEEE Trans. Multimedia*, vol. 12, no. 5, pp. 358–371, Aug.. 2010.

[10] M. I. Mandel, R. J. Weiss, and D. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 382–394, Feb. 2010.

[11] H. Sawada, S. Araki, and S. Makino, "A two-stage frequency-domain blind source separation method for underdetermined convolutive mixtures," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, New Paltz, NY, USA, 2007, pp. 139–142.

[12] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.

[13] P. D. O'Grady and B. A. Pearlmutter, "Soft-LOST: EM on a mixture of oriented lines," in *Proc. ICA 2004, ser. Lecture Notes in Computer Science, Springer-Verlag*, 2004, pp. 430–436.

[14] A. Alinaghi, W. Wang, and P. J. B. Jackson, "Integrating binaural cues and blind source separation method for separating reverberant speech mixtures," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2011, pp. 209–212.

[15] 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, Aug. 1987.

[16] S. M. Naqvi, M. Yu, and J. A. Chambers, "A multimodal approach to blind source separation of moving sources," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 5, pp. 895–910, Aug. 2010.

[17] 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.

[18] B. Ristic, S. Arulampalam, and N. Gordon, *Beyond the Kalman Filter: Particle Filter for Tracking Applications*. Norwood, MA: Artech House, 2004.

[19] L. Mihaylova and A. Carmi, "Particle algorithms for filtering in high dimensional state spaces: A case study in group object tracking," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2011, pp. 5932–5935.

[20] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Trans. Signal Process.*, vol. 50, no. 2, pp. 174–188, Feb. 2002.

[21] Y. Bar-Shalom and X. Li, *Estimation and Tracking: Principles, Techniques and Software*. Norwood, MA: Artech House, 1993, ",".

[22] P. Brasnett, L. Mihaylova, D. Bull, and N. Canagarajah, "Sequential Monte Carlo tracking by fusing multiple cues in video sequences," *Image Vis. Comput.*, vol. 25, no. 8, pp. 1217–1227, 2007.

[23] A. Rehman, S. M. Naqvi, W. Wang, R. Phan, and J. A. Chambers, "MCMC-PF based multiple head tracking in a room environment," in *Proc. 4th UK Computer Vision Student Workshop (BMVW)*, 2012.

[24] X. Xu and B. Li, "Head tracking using particle filter with intensity gradient and color histogram," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2005, pp. 888–891.

[25] 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.

[26] X. Birchfield, "An elliptical head tracker," in *Proc. 31st Asilomar Conf. Signals, Syst., Comput.*, 1997, vol. 2, pp. 1710–1714.

[27] A. Bhattacharayya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bull. Calcutta Math. Soc.*, vol. 35, pp. 99–110, 1943.

[28] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ. Press, 2001.

[29] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Part IV, Optimum Array Processing*. New York, NY, U.K.: Wiley, 2002.

[30] G. Lathoud, J. M. Odobez, and D. G. Perez, "AV16.3: An audio-visual corpus for speaker localization and tracking," in *Proc. MLMI'04 Workshop*, 2004, pp. 182–195.

[31] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic phonetic continuous speech corpus CDROM, 1993," [Online]. Available: http://www.ldc.upenn.edu/Catalog/LDC93S1W.html

[32] B. Shinn-Cunningham, N. Kopco, and T. Martin, "Localizing nearby sound sources in a classroom: Binaural room impulse responses," *J. Acoust. Soc. Amer.*, vol. 117, no. 5, pp. 3100–3115, 2005.

[33] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, 1979.

[34] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

[35] Y. Liang, S. M. Naqvi, and J. A. Chambers, "Audio video based fast fixed-point independent vector analysis for multisource separation in a room environment," *EURASIP J. Adv. Signal Process.*, 2012, 2012:183.

[36] H. Shabani and M. H. Kahaei, "Missing feature mask generation in BSS outputs using pitch frequency," in *Proc. 17th Int. Conf. Digital Signal Process.*, Corfu, Greece, 2011.

[37] A. U.-Rehman, S. M. Naqvi, R. Phan, and J. A. Chambers, "Multi-speaker direction of arrival tracking for multimodal source separation of moving sources," in *Proc. Sensor Signal Process. for Defence (SSPD '11)*, 2011, pp. 1–5.

[38] S. M. Naqvi, M. Yu, and J. A. Chambers, "Multimodal blind source separation for moving sources based on robust beamforming," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2011, pp. 241–244.

[39] K. Lebart, J. M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acust. United with Acust.*, vol. 87, no. 3, pp. 359–366, 2001.

[40] M. Jeub, M. Schäfer, T. Esch, and P. Vary, "Model-based dereverberation preserving binaural cues," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1732–1745, Sep. 2010.

Muhammad Salman Khan (S'06) received the B.S. (Hons.) and the M.S. degrees in electrical engineering from the University of Engineering and Technology (UET), Peshawar, Pakistan, in 2007, and the George Washington University, Washington, D.C., USA, in 2010, respectively. He worked as an Operation Engineer with the Pakistan Telecommunication Company Limited (PTCL) for three months in 2008 before proceeding to the US to commence his M.S. studies. He also served as a Lecturer with the Department of Electrical Engineering, UET, Peshawar, Pakistan, for eight months in 2010. He is currently pursuing the Ph.D. degree under the supervision of Prof. Jonathon Chambers within the Advanced Signal Processing Group at Loughborough University, Leicestershire, U.K.

His research interests include sound source separation, auditory-inspired audio processing, time-frequency masking, dereverberation, and computational auditory scene analysis. He is a student member of the IEEE and a member of the IEEE Signal Processing Society.

Syed Mohsen Naqvi (S'07–M'09) received his B.Eng. degree in industrial electronics engineering from IIEE/NED University of Engineering and Technology, Karachi, Pakistan, in 2001 and his Ph.D. degree in signal processing from Loughborough University, Leicestershire, U.K., in 2009. Before his postgraduate studies in U.K., he worked in research and development in Pakistan from January 2002 to September 2005. At present he is working as a Lecturer in Image and Video Processing in School of Electronic, Electrical and Systems Engineering, Loughborough University, Leicestershire, U.K., prior to this faculty position, from July 2009 to September 2012, he worked as a Postdoctoral Research Associate on the Engineering and Physical Sciences Research Council (EPSRC) of the U.K. funded projects.

Dr. Naqvi has authored or co-authored around 45 research outputs with main focus on his research area of multimodal (audio-visual) speech processing and his research interests include nonlinear filtering, data fusion and multi-target tracking. He is a member of the IEEE and the IEEE Signal Processing Society. He is a TPC member of the 16th International Conference on Information Fusion 2013.

Ata-ur-Rehman (S'13) received the B.Eng. degree in electronic engineering from Air University, Islamabad, Pakistan in 2006. He then joined the University of Engineering and Technology, Lahore, Pakistan, as a Lab Engineer in 2007. He moved to Loughborough University, U.K., in 2009 where he received his M.Sc. degree with distinction in Digital Communication Systems in 2010. Since 2010 he has been pursuing the Ph.D. degree in Loughborough University. His main area of research is multi-target tracking with applications in audio source separation.

Wenwu Wang (M'02–SM'11) was born in Anhui, China, in 1974. He received the B.Sc. degree in automatic control in 1997, the M.E. degree in control science and control engineering in 2000, and the Ph.D. degree in navigation guidance and control in 2002, all from Harbin Engineering University, Harbin, China. He then joined King's College, London, U.K., in May 2002, as a postdoctoral research associate and transferred to Cardiff University, Cardiff, U.K., in January 2004, where he worked in the area of blind signal processing. In May 2005, he joined the Tao Group, Ltd., Reading, U.K., as a DSP engineer working on algorithm design and implementation for real-time and embedded audio and visual systems. In September 2006, he joined Creative Labs, Ltd., Egham, U.K., as an R&D engineer, working on 3D spatial audio for mobile devices. Since May 2007, he has been with the University of Surrey, Guildford, U.K., where he is currently a Senior Lecturer, researching and teaching in signal processing. He is also a member of the MoD University Defense Research Collaboration (UDRC) in Signal Processing and the BBC Audio Research Partnership.

His current research interests include blind signal processing, sparse signal processing, audio-visual signal processing, machine learning and perception, machine audition (listening), and statistical anomaly detection. He has (co-)authored over 100 journal and conference papers in these areas, as well as a book entitled Machine Audition: Principles, Algorithms and Systems by IGI Global. He and his team of researchers have won the Best Solution Award on the DSTL Challenge Workshop for the signal processing challenge "under-sampled signal recognition" in 2012, and the Best Student Paper Award nomination on the 9th International Conference on Latent Variable Analysis and Signal Separation in 2010, the Hot Paper (feature article) on the Wiley/IEEE worldwide advert for publications in signal and image processing in 2008, and the IEEE Signal Processing Society Travel Grant in 2013.

Jonathon Chambers (S'83–M'90–SM'98–F'11) received the Ph.D. degree in signal processing from the Imperial College of Science, Technology and Medicine (Imperial College London), London, U.K., in 1990. From 1991 to 1994, he was a Research Scientist with Schlumberger Cambridge Research Center, Cambridge, U.K. In 1994, he returned to Imperial College London, as a Lecturer in signal processing and was promoted as a Reader (Associate Professor) in 1998. From 2001 to 2004, he was the Director of the Centre for Digital Signal Processing and a Professor of signal processing with the Division of Engineering, King's College London, London. From 2004 to 2007, he was a Cardiff Professorial Research Fellow with the School of Engineering, Cardiff University, Wales, U.K. In 2007, he joined the Department of Electronic and Electrical Engineering, Loughborough University, Loughborough, U.K., where he heads the Advanced Signal Processing Group.

He is a co-author of the books Recurrent Neural Networks for Prediction: Learning Algorithms, Architectures and Stability (Wiley, 2001) and EEG Signal Processing (Wiley, 2007). He has advised more than 50 researchers through to Ph.D. graduation and published more than 400 conference proceedings and journal articles, many of which are in IEEE journals. His research interests include adaptive and blind signal processing and their applications. Prof. Chambers is a Fellow of the Royal Academy of Engineering, U.K., and the Institution of Electrical Engineers (IEE). He was the Technical Program Chair of the 15th International Conference on Digital Signal Processing (DSP 2007) and the 2009 IEEE Workshop on Statistical Signal Processing, both held in Cardiff, U.K., and a Technical Program Cochair for the 36th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011), Prague, Czech Republic. He is the recipient of the first QinetiQ Visiting Fellowship in 2007 "for his outstanding contributions to adaptive signal processing and his contributions to QinetiQ" as a result of his successful industrial collaboration with the international defense systems company QinetiQ.

Prof. Chambers has served on the IEEE Signal Processing Theory and Methods Technical Committee for six years, the IEEE Signal Processing Society Awards Board for three years, and is currently a member of the IEEE Signal Processing Society Conference Board and the European Signal Processing Society Best Paper Awards Selection Panel. He has also served as an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING for three terms over the periods 1997–1999, 2004–2007 and 2011—(and is currently a Senior Area Editor).