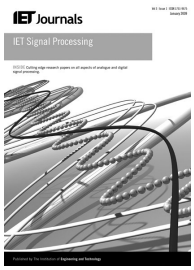
Techset Composition Ltd, Salisbury
Doc: {IEE}SPR/Articles/Pagination/SPRSI20110124.3d

www.ietdl.org

Published in IET Signal Processing
Received on 5th April 2011
Revised on 24th August 2011
doi: 10.1049/iet-spr.2011.0124



ISSN 1751-9675

Multimodal (audio–visual) source separation exploiting multi-speaker tracking, robust beamforming and time–frequency masking

S. Mohsen Naqvi[1] *W. Wang*[2] *M. Salman Khan*[1] *M. Barnard*[2] *J.A. Chambers*[1]

[1]*Advanced Signal Processing Group, School of Electronic, Electrical and Systems Engineering, Loughborough University, Leicestershire LE11 3TU, UK*
[2]*The Centre for Vision, Speech and Signal Processing, Department of Electronic Engineering, University of Surrey, Guildford, GU2 7XH, UK*
E-mail: s.m.r.naqvi@lboro.ac.uk

Q1 **Abstract:** A novel audio–visual approach is proposed to solve the problem of source separation of physically moving and stationary sources which exploits a circular microphone array, multiple video cameras, robust spatial beamforming and time–frequency masking. The challenge of separating moving sources is that the mixing filters are time varying; as such the unmixing filters should also be time varying but these are difficult to determine from only audio measurements. Even for physically stationary sources measured in reverberant room environments, for example with reverberation time (RT) greater than 300 ms, the performance of audio-only blind source separation methods is limited. Therefore in the proposed approach, the visual modality is used to facilitate the separation for both stationary and moving sources. The movement of the sources is detected by a three-dimensional (3D) tracker based on a Markov Chain Monte Carlo particle filter. The audio separation is performed by exploiting a robust least squares frequency invariant data-independent beamformer which is shown to be particularly suitable for room environments with RT greater than 300 ms. The uncertainties in source localisation and direction of arrival information obtained from the 3D video-based tracker are controlled by using a convex optimisation approach in the beamformer design. In the final stage, the separated audio sources are further enhanced by applying a binary time–frequency masking technique as a post-filtering process in the cepstral domain. Experimental results show that using the visual modality, the proposed algorithm cannot only achieve performance better than conventional frequency-domain source separations algorithms, but also provide acceptable separation performance for moving sources.

1 Introduction

The main objective of machine-based speech separation is to mimic the ability of a human to separate multiple sound sources from their sound mixtures, that is, to provide a solution to the so-called cocktail party problem. This problem was coined by Colin Cherry in 1953 [1], who first asked the question: 'How do we [humans] recognize what one person is saying when others are speaking at the same time?'. Despite being studied extensively, it remains a scientific challenge as well as an active research area. A main stream of effort over the past decade in the signal processing community has been to address the problem under the framework of convolutive blind source separation (CBSS) where the sound recordings are modelled as linear convolutive mixtures of the unknown speech sources. The solutions to CBSS were formulated initially in the time domain, which was soon shown to be computationally expensive for a real room environment as a large sample length, typically on the order of thousands of samples, is needed to represent the room impulse responses [2, 3]. To address this problem, the frequency domain approaches were then proposed, in which the CBSS is simplified to complex-valued instantaneous BSS problems at each frequency bin, subject to the two indeterminacies, that is, scaling and permutation, which are inherent to instantaneous BSS. These ambiguities are more severe in the frequency domain, therefore many methods have been developed to solve them, see for example [4, 5]. However, the state-of-the-art algorithms still commonly suffer in the following two practical situations, namely, for the highly reverberant environment, and when multiple moving sources are present. In both cases, most existing methods [6–12] have poor separation performance for the most part because of the data length limitations.

The algorithms discussed above are all unimodal, that is, operating only in the audio domain. However, as is widely accepted, both speech production and perception are inherently multimodal processes. On the one hand, the production of speech is usually coupled with the visual movement of the mouth and facial muscles. On the other hand, looking at the lip movement of a speaker (i.e. lip reading) is helpful for listeners to understand what has been said in a conversation, in particular, when multiple

© The Institution of Engineering and Technology 2012

competing conversations and background noise are present simultaneously, as shown in an early work in [13]. The well-known McGurk effect [14] also confirms that visual articulatory information is integrated into the human speech perception process automatically and unconsciously [15]. As also suggested by Cherry [16], combining the multimodal information from different sensory measurements would be the best way to address the machine cocktail party problem. Recent studies in this direction, that is, integrating the visual information into audio-only speech source separation systems, are emerging as an exciting new area in signal processing, that is, multimodal speech separation [17–23].

Visual information is complementary to the audio modality, as a result it has great potential for overcoming the limitations of existing CBSS algorithms especially in adverse situations (such as rooms with large reverberation time (RT) and moving source scenarios) [19, 20, 24]. The visual information has, moreover, been shown to be useful for the cancellation of the two ambiguities of BSS algorithms [22, 25, 26].

Most existing BSS algorithms assume that the sources are physically stationary, that is, the mixing filters are fixed. All these algorithms are based on statistical information extracted from the received mixed audio data and are generally unable to operate for separation of moving sources because of data length limitations, that is, the number of samples available at each frequency bin is not sufficient for the algorithms to converge [27]. Therefore new BSS methods for moving sources are required to solve the cocktail party problem in practice. Only a few papers have been presented in this area [20, 28–33]. In our recent work [20], a visual tracker was implemented for DOA information and a simple beamformer in linear array configuration was used to enhance the signal from one source direction and to reduce the energy received from another source direction in a low reverberant environment.

For rooms with higher RT, the performance of conventional CBSS algorithms is also limited. Increasing the frame length of the short time Fourier transform (STFT) to compensate for the length of the mixing filters may violate the independence assumption necessary in BSS [34]. Moreover, the separated speech signals with CBSS for higher RTs are perceptually poor, because the reverberations are not well suppressed. On the other hand, beamforming accepts the direct path and also suppresses the later reflections that arrive from directions where the response of the beamformer is low. As such beamforming has much potential for source separation in rooms with large RTs.

In computational auditory scene analysis (CASA), a recent technique which uses a binary time–frequency mask known as an ideal binary mask (IBM), has shown promising results in interference removal and improving intelligibility of the target speech [35, 36]. Originally, the IBM technique was proposed as a performance benchmark of a CASA system [37]. This IBM technique also has the potential to be used as post-filtering to provide a significant improvement in the separation obtained from the robust least squares frequency invariant data-independent (RLSFIDI) beamformer.

In this paper, a multimodal source separation approach is therefore proposed by utilising not only received linearly mixed audio 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 velocities, that is, stationary or moving, of the speakers. The RLSFIDI beamformer is employed with linear and circular array configurations for multiple speakers and realistic three-dimensional (3D) scenarios for physically moving and stationary sources. The velocity information of each speaker and DOA information to the microphone array are obtained from a 3D visual tracker based on the Markov Chain Monte Carlo particle filter (MCMC-PF) from our work in [20]. In the RLSFIDI beamformer, we exploit 16 microphones to provide sufficient degrees of freedom to achieve more effective interference removal. To control the uncertainties in source localisation and direction of arrival information, constraints to obtain wider main lobe for the source of interest (SOI) and to better block the interference are exploited in the beamformer design. The white noise gain (WNG) constraint is also imposed which controls robustness against the errors because of mismatch between sensor characteristics [38]. Although the RLSFIDI beamformer provides a good separation performance for physically moving sources in a low reverberation environment, for higher reverberance situations the beamforming approach can only reduce the reflections from certain directions. The separation performance is therefore further enhanced using the IBM technique as a post-filtering process stage. The RLSFIDI beamformer is also shown to provide better separation than state-of-the-art CBSS methods for physically stationary sources within room environments with RT $>$ 300 ms. The proposed approach can be divided into two parts: 3D multi-speaker visual tracking and robust beamformer plus time–frequency masking-based source separation. The schematic diagram of the system is shown in Fig. 1.

The remainder of the paper is organised as follows. Section 2 provides the video tracking. Section 3 presents audio source separation including frequency invariant data independent beamformer design for a circular array configuration in a 3D room environment and the method used for post-processing, that is, the IBM technique with cepstral smoothing. Experimental results are discussed in Section 4. In Section 5, we conclude the paper.

2 Video tracking

The 3D visual tracker is based on the MCMC-PF, which results in high sampling efficiency [39, 40]. The detect before track approach is used.

Detection: Video localisation is based on face and head detection, an off-the-shelf, state-of-the-art, Viola–Jones face detector [41] is used. It is highlighted that the area of detection is a discipline in its own rights and here we simply exploit recent results from this field to provide geometric information to facilitate an audio–visual approach to CBSS. Audio localisation for multiple simultaneously active speakers in a room environment can fail. Localisation 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 [42]. Video localisation is not always effective, especially when the face of a human being is not visible to at least two cameras because of some obstacles, the environment is cluttered, camera angle is wide or illumination conditions are varying. A combination of audio–visual modalities with multiple camera integration is the most suitable choice for source localisation and is our future work. In this paper, for video localisation, we used face and head detection because there is no guarantee that video cameras can capture a frontal view of the human, therefore using face and head detection has an advantage for practical applications [43]. In order to approximate the 3D position of the speakers in each sequence we use at least two out of four calibrated colour

© The Institution of Engineering and Technology 2012

doi: 10.1049/iet-spr.2011.0124

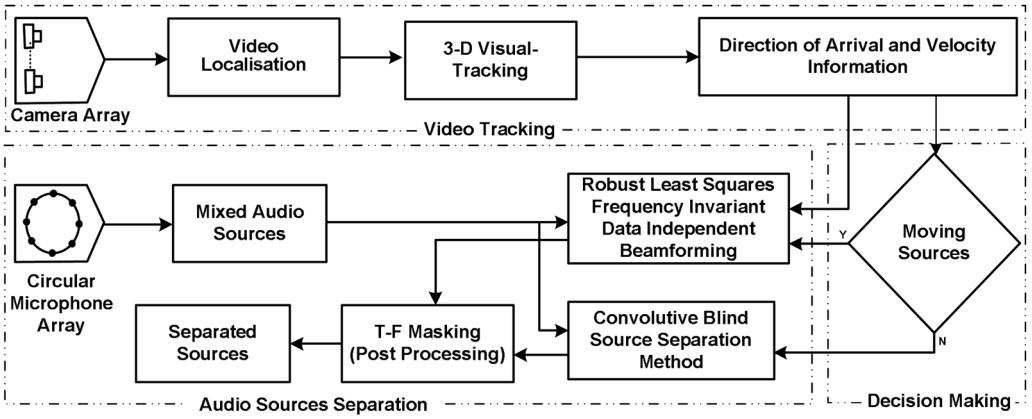


Fig. 1 *System block diagram: video localisation is based on the face and head detection*

The 3D location of each speaker is approximated after processing the 2D image information and obtained from at least two synchronised colour video cameras through calibration parameters and an optimisation method. The approximate 3D locations are fed to the visual-tracker based on an MCMC-PF to estimate the 3D real-world positions. The position of the microphone array and the output of the visual tracker are used to calculate the direction of arrival and velocity information of each speaker. Based on the velocity information of the speakers the audio mixtures obtained from the linear or circular microphone array configurations are separated either by an RLSFIDI beamformer or by a CBSS algorithm

video cameras [44], synchronised with the external hardware trigger module and an optimisation method [45].

Tracking: The approximate 3D location of each speaker is estimated by using the Bayesian multi-speaker state space approach. The 3D multispeaker observation is defined as $\boldsymbol{Z}_{1:k} = \boldsymbol{Z}_{1,\ 1:k}, \ldots, \boldsymbol{Z}_{n,1:k}$, where $\boldsymbol{Z}_{i,1:k}$ represents the observations of speaker i over the discrete-time interval 1, 2, …, k, written as 1:k as in MATLAB notation, and the multispeaker state configuration is defined as $\boldsymbol{U}_{1:k} = \boldsymbol{U}_{1,1:k}, \ldots, \boldsymbol{U}_{n,1:k}$. The filtering distribution of states given observations $p(\boldsymbol{U}_k|\boldsymbol{Z}_{1:k})$ is recursively approximated using an MCMC-PF, which provides high sampling efficiency. The detailed description of the three important items of the probabilistic multi-speaker 3D visual tracker, the state model, the measurement model and the MCMC-sampling mechanisms are provided in our work [20].

The implementation steps for the 3D visual tracker based on MCMC-PF algorithm are as follows:

1. MCMC-PF algorithm takes in 3D approximated position $\boldsymbol{Z}_{1,k}$ of each speaker i at each state k, which is obtained from face and head detection in the images from at least two synchronised colour video cameras and an optimisation method.
2. *Initialise the MCMC sampler:* At time k predict the state of each speaker i for N_p particles, that is, $\{\boldsymbol{U}^n_{i,k},\}^{N_\mathrm{p}}_{n=1}$ from the particle set at time $k-1$, that is, $\{\boldsymbol{U}^n_{i,k-1}\}^{N_\mathrm{p}}_{n=1}$ based on the factorised dynamic model $\Pi_i p(\boldsymbol{U}_{i,k}|\boldsymbol{U}_{i,k-1})$.
3. $B + N_\mathrm{p}$ *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 assumed to move.
- Sample a new state $\boldsymbol{U}^*_{i,k}$ for only speaker i from the single speaker proposal density $Q(\boldsymbol{U}^*_{i,k}|\boldsymbol{U}_{i,k})$ [20].
- Compute the acceptance ratio for the evaluation of likelihood for only speaker i [20]

$$\alpha = \min\left\{1, \frac{p(\boldsymbol{Z}_{i,k}|\boldsymbol{U}^*_{i,k})Q(\boldsymbol{U}_{i,k}|\boldsymbol{U}^*_{i,k})}{p(\boldsymbol{Z}_{i,k}|\boldsymbol{U}_{i,k})Q(\boldsymbol{U}^*_{i,k}|\boldsymbol{U}_{i,k})}\right\} \tag{1}$$

- Draw μ from uniform distribution.
- If $\alpha > \mu$ then accept the move for speaker i and change the $\boldsymbol{U}^*_{i,k}$ into $\boldsymbol{U}_k$. Otherwise, reject the move, do not change $\boldsymbol{U}_k$ and copy to the new sample set.

4. Discard the first B samples to form the particle set, $\{\boldsymbol{U}^n_k\}^{N_\mathrm{p}}_{n=1}$, at time step k. The output of the 3D tracker at each state k is the mean estimate for each speaker.
5. After estimating the 3D position of each speaker i the elevation (θ_i) and azimuth (ϕ_i) angles of arrival to the centre of the microphone array are calculated.

$$r_i = \sqrt{(u_{x_i} - u'_{x_m})^2 + (u_{y_i} - u'_{y_m})^2 + (u_{z_i} - u'_{z_m})^2} \tag{2}$$

$$\theta_i = \tan^{-1}\left(\frac{u_{y_i} - u'_{y_m}}{u_{x_i} - u'_{x_m}}\right) \tag{3}$$

$$\phi_i = \mathrm{Sin}^{-1}\left(\frac{u_{y_i} - u'_{y_m}}{r_i\,\mathrm{Sin}(\theta_i)}\right) \tag{4}$$

where u_{x_i}, u_{y_i} and u_{z_i} are the 3D positions of the moving speaker i, whereas u'_{x_m}, u'_{y_m} and u'_{z_m} are Cartesian coordinates of the centre of microphone array.

The above DOA information is fed to the RLSFIDI beamformer. The details of the CBSS process, the RLSFIDI beamformer and T–F masking are presented in the following section.

3 Audio source separation

3.1 Convolutive blind source separation

In CBSS, for M audio sources recorded by N microphones, the convolutive audio mixtures obtained can be described mathematically as

$$x_i(n) = \sum_{j=1}^{M}\sum_{p=1}^{P} h_{ij}(p)s_j(n-p+1) \tag{5}$$

where s_j is the source signal from a source $j = 1, \ldots, x_i$ is the received signal by microphone $i = 1, \ldots, N$ and $h_{ij}(p)$, $p = 1, \ldots, P$, is the p-tap coefficient of the impulse response from

© The Institution of Engineering and Technology 2012

source j to microphone i and n, is the discrete time index. In this work, we assume $N \geq M$.

In time domain CBSS, the sources are estimated using a set of unmixing filters such that

$$y_j(n) = \sum_{i=1}^{N} \sum_{q=1}^{Q} w_{ji}(q) x_i(n - q + 1) \tag{6}$$

where $w_{ji}(q)$, $q = 1, \ldots, Q$, is the q-tap weight from microphone i to source j.

The CBSS problem in the time domain can be converted to multiple complex-valued instantaneous problems in the frequency domain by using a T-point windowed STFT. The time domain signals $x_i(n)$, are converted into time–frequency domain signals $x_i(\omega, k)$, where ω and k are respectively, frequency and time frame indices. The N observed mixed signals can bedescribed as a vector in the time–frequency domain as

$$\boldsymbol{x}(\omega, k) = \boldsymbol{H}(\omega)\boldsymbol{s}(\omega, k) \tag{7}$$

where $\boldsymbol{x}(\omega, k)$ is an $N \times 1$ observation column vector for frequency bin ω, $\boldsymbol{H}(\omega)$ is an $N \times M$ mixing matrix, $\boldsymbol{s}(\omega, k)$ is an $M \times 1$ speech sources vector and the source separation can be described as

$$\boldsymbol{y}(\omega, k) = \boldsymbol{W}(\omega)\boldsymbol{x}(\omega, k) \tag{8}$$

where $\boldsymbol{W}(\omega)$ is $M \times N$ separation matrix. By applying an inverse STFT (ISTFT), $\boldsymbol{y}(\omega, k)$ can be converted back to the time domain audio signals as

$$y(n)=\text{ISTFT}(\boldsymbol{y}(\omega, k)) \tag{9}$$

The audio signals are separated with the help of the above-mentioned visual information obtained from the 3D tracker. Therefore RLSFIDI beamformer design for linear and circular array configurations in a 3D room environment is explained in the following section.

3.2 RLSFIDI beamformer

The least squares approach is a suitable choice for data-independent beamformer design [46], by assuming the over-determined case $N > M$, which provides greater degrees of freedom, we obtain the over-determined least squares problem for the beamformer design for one of the sources as

$$\min_{\boldsymbol{w}(\omega)} ||\boldsymbol{H}^{\mathrm{T}}(\omega)\boldsymbol{w}(\omega) - \boldsymbol{r}_{\mathrm{d}}(\omega)||_2^2 \tag{10}$$

where $\boldsymbol{r}_{\mathrm{d}}(\omega)$ is an $M \times 1$ desired response vector and can be designed from a 1D window (e.g. the Dolph–Chebyshev or Kaiser windows), $\boldsymbol{w}^{\mathrm{T}}(\omega)$ is one of the beamformer weight vectors which corresponds to one row vector of $\boldsymbol{W}(\omega)$ in (8), and $(\cdot)^{\mathrm{T}}$ and $\|\cdot\|_2$ denote respectively the vector transpose operator and the Euclidean norm.

A frequency invariant beamformer design can be obtained by assuming the same coefficients for all frequency bins, that is $\boldsymbol{r}_{\mathrm{d}}(\omega) = \boldsymbol{r}_{\mathrm{d}}$ [47]. If the wavelengths of the low frequencies of the sources are greater than twice the spacing between the microphones then this design leads to spatially white noise [38]. In unimodal (audio only) CBSS systems there are no priori assumptions on the source statistics of the mixing system. On the other hand, in the audio–visual approach the video system can capture the positions of the speakers and the directions they face. Therefore the mixing filter is formulated as $\boldsymbol{H}(\omega) = [\boldsymbol{d}(\omega, \theta_1, \phi_1), \ldots, \boldsymbol{d}(\omega, \theta_M, \phi_M)]$, and is based on the visual information, that is, the DOA from the 3D visual tracker, where $\boldsymbol{d}(\cdot)$ denotes the beamformer response vector.

An N-sensor circular array with a radius of R and a target speech signal having DOA information (θ, ϕ), where θ and ϕ are the elevation and azimuth angles respectively, is shown in Fig. 2. The sensors are equally spaced around the circumference, and their 3D positions, which are calculated from the array configuration, are provided in the matrix form as

$$\boldsymbol{U}' = \begin{bmatrix} u'_{x_1} & u'_{y_1} & u'_{z_1} \\ \vdots & \vdots & \vdots \\ u'_{x_N} & u'_{y_N} & u'_{z_N} \end{bmatrix} \tag{11}$$

where the Cartesian coordinates of the nth sensor (microphone) are in the nth row of matrix $\boldsymbol{U}'$. It is highlighted that the matrices $\boldsymbol{U}$ and $\boldsymbol{U}'$ contains the 3D positions of the speakers and microphone array, respectively.

The beamformer response $\boldsymbol{d}(\omega, \theta_i, \phi_i)$ for frequency bin ω and for SOI $i = 1, \ldots, M$, can be derived [48] as

$$\boldsymbol{d}(\omega, \theta_i, \phi_i) = \begin{bmatrix} \exp(-j\kappa(\sin(\theta_i).\cos(\phi_i).u'_{x_1} + \sin(\theta_i) \\ \sin(\phi_i).u'_{y_1} + \cos(\theta_i).u'_{z_1})) \\ \vdots \\ \exp(-j\kappa(\sin(\theta_i).\cos(\phi_i).u'_{x_N} + \sin(\theta_i). \\ \sin(\phi_i).u'_{y_N} + \cos(\theta_i).u'_{z_N})) \end{bmatrix} \tag{12}$$

where $\kappa = \omega/c$ and c is the speed of sound in air at room temperature. It is highlighted that (11) and (12) are valid for different microphone array configurations in a 3D realistic environment, and we also used a linear array configuration in this paper.

To design the beam pattern, which allow the SOI and to better block the interference, in the least squares problem in

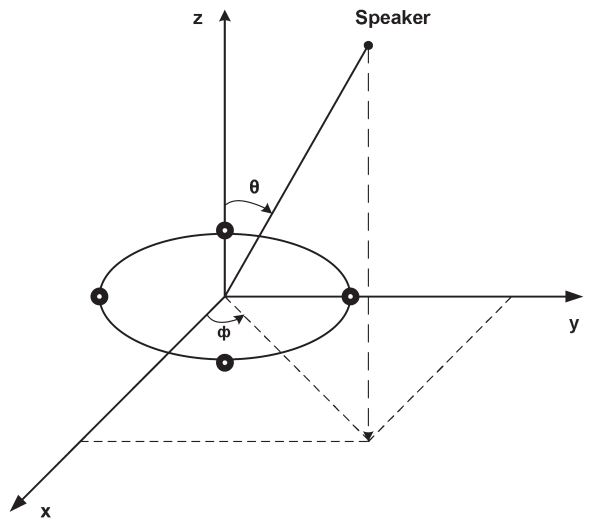


Fig. 2 *Circular array configuration for a 3D realistic environment*

© The Institution of Engineering and Technology 2012

doi: 10.1049/iet-spr.2011.0124

(10), the following constraints are used

$$|\boldsymbol{w}^{\mathrm{H}}(\omega)\boldsymbol{d}(\omega, \theta_i + \Delta\theta, \phi_i + \Delta\phi)| = 1$$
$$|\boldsymbol{w}^{\mathrm{H}}(\omega)\boldsymbol{d}(\omega, \theta_j + \Delta\theta, \phi_j + \Delta\phi)| < \varepsilon \quad \forall\omega \tag{13}$$

where θ_i, ϕ_i and θ_j, ϕ_j, $j = 1, \ldots, M$ except i, are respectively, the angles of arrival of the SOI and interference, and $\Delta\theta$ and $\Delta\phi$ have angular ranges defined by $\alpha_1 \leq \Delta\theta \leq \alpha_2$ and $\alpha_3 \leq \Delta\phi \leq \alpha_4$, where α_1, α_3 and α_2, α_4 are lower and upper limits respectively, and ε is the bound for interference. To control the uncertainties in source localisation and direction of arrival information the angular ranges $\theta_i + \Delta\theta$, $\phi_i + \Delta\phi$, with $\Delta\theta \in [\alpha_1, \alpha_2]$ and $\Delta\phi \in [\alpha_3, \alpha_4]$, is divided into discrete values which thereby provide the wider main lobe for the SOI and wider attenuation beam pattern to block the interferences.

The WNG is a measure of the robustness of a beamformer and a robust superdirectional beamformer can be designed by constraining the WNG. Superdirective beamformers are extremely sensitive to small errors in the sensor array characteristics and to spatially white noise. The errors because of array characteristics are nearly uncorrelated from sensor to sensor and affect the beamformer in a manner similar to spatially white noise. The WNG is also controlled in this paper by adding the following constraint

$$\boldsymbol{w}^{\mathrm{H}}(\omega)\boldsymbol{w}(\omega) \leq \frac{1}{\gamma} \quad \forall\omega \tag{14}$$

where γ is the bound for the WNG.

The constraints in (13) for each discrete pair of elevation and azimuth angles, the respective constraint for WNG in (14) are convex [38]. In addition, the unconstrained least square problem in (10) is a convex function, therefore convex optimisation [49] is used to calculate the weight vector $\boldsymbol{w}(\omega)$ for each frequency bin ω. The RLSFIDI beamformer design for each discrete set of angles may vary from the delay sum beamformer to a desired highly sensitive supereffective beamformer, depending on the bound for WNG $\gamma \ll 1$. This flexibility can be used to adapt any given prior knowledge on microphone mismatch, microphone self-noise and positioning errors [38].

Finally, $\boldsymbol{W}(\omega) = [\boldsymbol{w}_1(\omega), \ldots, \boldsymbol{w}_M(\omega)]^{\mathrm{T}}$ is placed in (8) to estimate the sources. As the scaling is not a major issue [7] and there is no permutation problem, the estimated sources are aligned for reconstruction in the time domain. These estimated sources are further enhanced by applying the time–frequency masking technique, discussed in the following section.

3.3 Combining the RLSFIDI beamformer and T–F masking

As mentioned above, the RLSFIDI beamformer passes the target signal from a certain direction (DOA obtained from video tracking) and suppresses interference and reflections, but the removal of interference is not perfect, therefore the IBM technique is used as a post-processing stage. The block diagram of combining the output of the RLSFIDI beamformer and T–F masking is shown in Fig. 3. The inherent scaling and permutation ambiguities are already mitigated, respectively, by normalising the sources weight vectors [20] and by using DOA in the RLSFIDI beamformer. Therefore the separated time domain speech signal $y_i(n)$ of speaker i is converted into the time–frequency T–F domain signals $y_i(\omega, k)$, where ω is a normalised frequency index. Using a T-point-windowed discrete STFT the spectrograms are obtained as

$$y_i(\omega, k) = \mathrm{STFT}(y_i(n)) \quad i = 1, \ldots, M \tag{15}$$

where k and ω, respectively, represent time and frequency bin indices.

From the above T–F units, binary masks are estimated by comparing the amplitudes of the spectrograms [35, 36] and in this paper we assume three audio sources, as previous work has only considered two sources. So that we can determine the estimated binary masks as (see (16)–(18))
where τ is a parameter to control how much of the interfering signals should be removed at each iteration [35, 36], and $\tau = 1$ is used in this paper. Then, each of the three binary masks is applied to the original mixtures (three mixtures are selected based on the geometric information obtained from 3D visual tracker) in the time–frequency domain to enhance the separated signals

$$y_i(\omega, k) = \mathrm{BM}_i(\omega, k)x_i(\omega, k) \quad i = 1, ., 3 \tag{19}$$

These speech signals are transformed in the time domain by applying an ISTFT.

This binary mask-based T–F technique considerably improves the separation performance of the RLSFI beamformer by reducing the interferences to a much lower level, which ultimately provides a good quality separated speech signal. However, an important problem with the binary masking is the introduction of errors in the estimation of the masks, that is, fluctuating musical noise [36]. To overcome this problem a cepstral smoothing technique [36, 50] is used.

3.3.1 Cepstral smoothing technique: In the cepstral smoothing, the estimated IBM is first transformed into the cepstral domain, and different smoothing levels, based on the speech production mechanism, are then applied to the transformed mask. The smoothed mask is further converted back to the spectral domain. In this method, the musical artifacts within the signals can be reduced. The broadband structure and pitch information of the speech signal are also well preserved without being noticeably affected by the smoothing operation [36]. The estimated masks in (16),

$$\mathrm{BM}_1(\omega, k) = \begin{cases} 1, & \text{if } |y_1(\omega, k)| > \tau|y_2(\omega, k)| \quad \& \quad |y_1(\omega, k)| > \tau|y_3(\omega, k)| \\ 0, & \text{otherwise} \quad \forall(\omega, k) \end{cases} \tag{16}$$

$$\mathrm{BM}_2(\omega, k) = \begin{cases} 1, & \text{if } |y_2(\omega, k)| > \tau|y_3(\omega, k)| \quad \& \quad |y_2(\omega, k)| > \tau|y_1(\omega, k)| \\ 0, & \text{otherwise} \quad \forall\ (\omega, k) \end{cases} \tag{17}$$

$$\mathrm{BM}_3(\omega, k) = \begin{cases} 1, & \text{if } |y_3(\omega, k)| > \tau|y_1(\omega, k)| \quad \& \quad |y_3(\omega, k)| > \tau|y_2(\omega, k)| \\ 0, & \text{otherwise} \quad \forall(\omega, k) \end{cases} \tag{18}$$

© The Institution of Engineering and Technology 2012

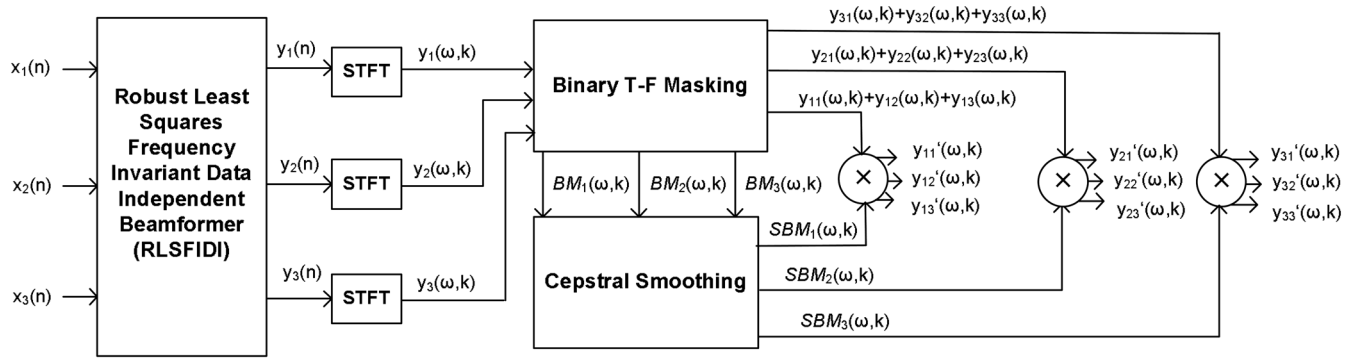


Fig. 3 *RLSFIDI beamformer with binary T–F masking as a post-processing technique provides estimates of the speech sources $y_i(n)$, where $i = 1, .., M$*

Separated speech signals are transformed to the T–F domain $y_i(\omega, k)$, using the STFT. Binary masks $\mathrm{BM}_i(\omega, k)$ are then estimated by comparing the energies of the individual T–F units of the source spectrograms. The cepstral smoothing stage follows that smooths the estimated binary masks and we obtain $\mathrm{SBM}_i(\omega, k)$. The smoothed binary masks $\mathrm{SBM}_i(\omega, k)$ are used to enhance the separated signals estimated by the RLSFIDI beamformer

(13) and (14) can be represented in the cepstral domain as

$$\mathrm{BM}_i^{\mathrm{c}}(l, k) = \mathrm{DFT}^{-1}\{\ln(\mathrm{BM}_i(\omega, k))|_{\omega=0,..,T-1}\} \quad i = 1, ..., 3 \tag{20}$$

where l is the quefrency bin index, DFT and ln denote the discrete Fourier transform and the natural logarithm operator respectively, T is the length of the DFT and after applying smoothing, the resultant smoothed mask is given as

$$\mathrm{BM}_i^{\mathrm{s}}(\omega, k) = \beta_l \mathrm{BM}_i^{\mathrm{s}}(l, k-1) + (1-\beta_l)\mathrm{BM}_i^{\mathrm{c}}(l, k) \tag{21}$$

where β_l controls the smoothing level and is selected according to different values of quefrency l

$$\beta_l = \begin{cases} \beta_{\mathrm{env}}, & \text{if} \quad l \in \{0, \ldots, l_{\mathrm{env}}\} \\ \beta_{\mathrm{pitch}}, & \text{if} \quad l = l_{\mathrm{pitch}} \\ \beta_{\mathrm{peak}}, & \text{if} \quad l \in \{(l_{\mathrm{env}}+1), \ldots, T\} \backslash l_{\mathrm{pitch}} \end{cases} \tag{22}$$

where l_{env} and β_{pitch} are respectively quefrency bin indices for the spectral envelope and the structure of the pitch harmonics in $\mathrm{BM}_i(\omega, k)$, and $0 \leq \beta_{\mathrm{env}} < \beta_{\mathrm{pitch}} < \beta_{\mathrm{peak}} \leq 1$. The symbol '\' excludes l_{pitch} from the quefrency range $(l_{\mathrm{env}}+1), \ldots, T$. The details of the principle for the range of β_l and the method to calculate β_{peak} are presented in [36]. The final smoothed version of the spectral mask is given as

$$\mathrm{SBM}_i(\omega, k) = \exp\left(\mathrm{DFT}\{\mathrm{BM}_i^{\mathrm{s}}(\omega, k)|_{l=0,\ldots,T-1}\}\right) \tag{23}$$

This smoothed mask is then applied to the segregated speech signals in (19) as follows

$$\bar{y}_i(\omega, k) = \mathrm{SBM}_i(\omega, k) y_i(\omega, k) \tag{24}$$

Finally, by applying an ISTFT, $\bar{y}_i(\omega, k)$ is converted back to the time domain audio signals. The experimental results based on objective and subjective evaluations are presented in the following section.

4 Experiments and results

Data Collection: The simulations are performed on audio–visual signals generated from a room geometry as illustrated in Fig. 4. Data were collected in a $4.6 \times 3.5 \times 2.5\ m^3$ smart office. Four calibrated colour video cameras ($C1$, $C2$, $C3$ and $C4$) were used to collect the video data. Video cameras were fully synchronised with an external hardware trigger module and frames were captured at 25 Hz with an image size of 640×480 pixels. For BSS evaluation, audio recordings of three speakers $M = 3$ were recorded at 8 KHz with linear and circular array configurations of 16 microphones, $N = 16$, equally spaced in line and around the circumference respectively. In the linear array configuration, the distance between the microphones was 4 cm and the radius of the circular array was $R = 0.2$ m. The other important variables were selected as: STFT length $T = 1024$ and 2048 and filter lengths were $Q = 512$ and 1024, the Hamming window is used with an overlap factor set to 0.75. The durations of the speech signals were 0.5 and 7 s, respectively, for moving and physically stationary sources, $\varepsilon = 0.1$, $\gamma = -10$ dB, for SOI $\alpha_1 = +5^\circ$ and $\alpha_2 = -5^\circ$, for interferences $\alpha_1 = +7^\circ$ and $\alpha_2 = -7^\circ$, speed of sound $c = 343$ m/s, $l_{\mathrm{env}} = 8$, $l_{\mathrm{low}} = 16$ and $l_{\mathrm{high}} = 120$, parameters for controlling the smoothing level were $\beta_{\mathrm{env}} = 0$, $\beta_{\mathrm{pitch}} = 0.4$, $\beta_{\mathrm{peack}} = 0.8$ and the room impulse duration RT60 = 130 ms (RT60 is the reverberation timeRT when the signal energy decays by 60 dB relative to its starting value). Speaker 2 was physically stationary and speakers 1 and 3 were moving. The same room dimensions, microphone locations and configuration, and selected speakers locations were used in the image method [51] to generate the audio data for

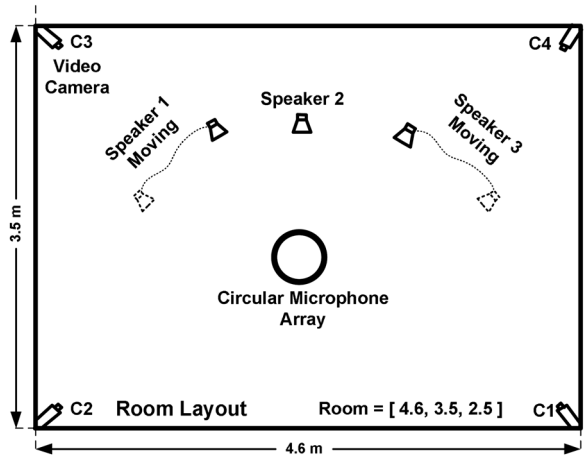


Fig. 4 *Room layout and audio–visual recording configuration*

© The Institution of Engineering and Technology 2012

doi: 10.1049/iet-spr.2011.0124

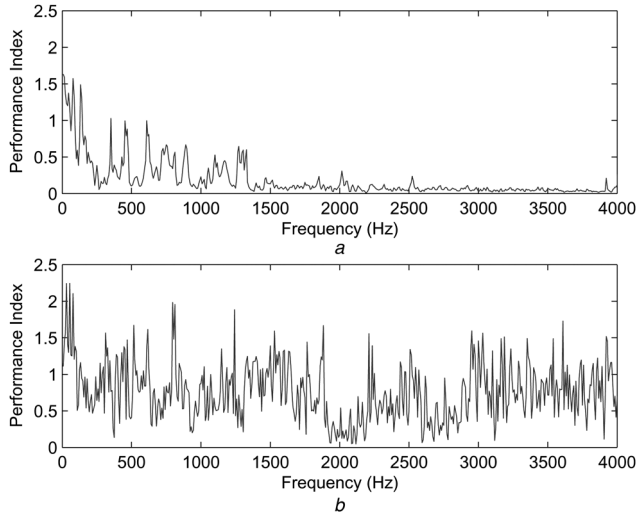


Fig. 5 *Performance index at each frequency bin for*

a RLSFIDI beamformer
b Original IVA method [54]
Length of the signals is 0.5 s. A lower PI refers to a superior method. The performance of the IVA method is poor because the CBSS algorithm cannot converge because of a limited number of samples in each frequency bin

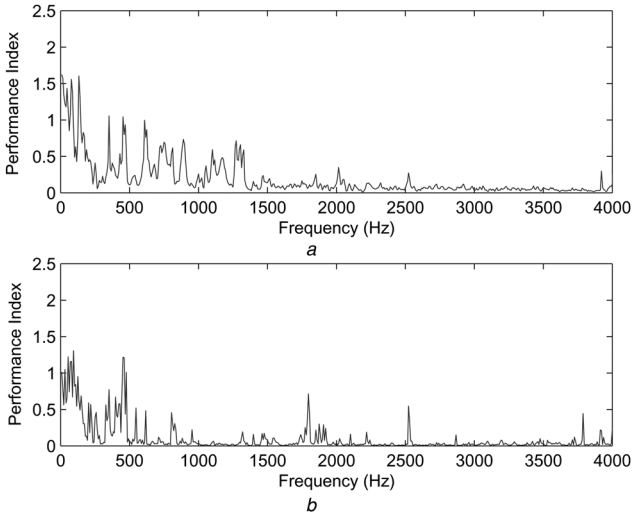


Fig. 6 *Performance index at each frequency bin for*

a RLSFIDI beamformer
b Original IVA method [54]
Length of the signals is 7 s. A lower PI refers to a superior method. The performance of the IVA method is better than RLSFIDI beamformer at RT60 = 130 ms

Table 1 Objective evaluation: ΔSINR, SDR, SIR and SAR for the RLSFIDI beamformer at RT60 = 130 ms, the length of the signals is 0.5 s[a]

ΔSINR	SDR_1	SDR_2	SDR_3	SIR_1	SIR_2	SIR_3	SAR_1	SAR_2	SAR_3
14.70	7.53	0.78	11.59	8.30	7.39	12.24	16.03	2.58	20.41

[a]Results are in decibels (dB)

Table 2 Objective evaluation: ΔSINR, SDR, SIR and SAR for the RLSFIDI beamformer and the original IVA method [54] at RT60 = 130 ms, the length of the signals is 7 s[a]

	ΔSINR	SDR_1	SDR_2	SDR_3	SIR_1	SIR_2	SIR_3	SAR_1	SAR_2	SAR_3
RLSFIDI beamformer	14.97	6.38	0.78	11.59	14.88	7.39	12.24	7.18	2.58	20.41
	15.03	7.42	0.53	11.81	8.63	7.00	12.87	14.13	2.43	18.66
IVA method	16.49	2.35	4.44	5.90	2.90	6.00	7.84	13.04	10.61	10.99
	16.35	2.40	4.51	6.10	3.30	6.06	8.14	13.32	10.69	11.07

[a]Results are in decibels (dB)

*RT*60 = 300 450 600 ms similar to our work [52]. The RT was controlled by varying the absorption coefficient of the walls.

Evaluation criteria: The objective evaluation of BSS includes performance index (PI) [20], signal-to-interference-noise ratio (SINR) and $\Delta\text{SINR} = \text{SINR}_o - \text{SINR}_i$, percentage of energy loss (PEL), percentage of noise residue (PNR) [35], signal-to-distortion ratio (SDR), signal-to-interference (SIR) ratio and signal-to-artifact ratio (SAR) [53]. The separation of the speech signals is evaluated subjectively by listening tests and mean opinion scores (MOS tests for voice are specified by ITU-T recommendation P.800) are also provided.

In the first simulation, the recorded mixtures of length = 0.5 s (which corresponds to the moving sources case) were separated by the original IVA method [54] and the RLSFIDI beamformer. The elevation angles from the 3D tracker for speakers 1, 2 and 3 were -81°, 90° and 80°, respectively. The azimuth angles for speakers 1, 2 and 3 were -45°, 80° and 45°, respectively. The DOA is passed to the RLSFIDI beamformer and the resulting performance indices are shown in Fig. 5*a*, which indicates good performance, that is, close to zero across the majority of the frequencies. The other objective evaluations are shown in Table 1. This separation quality was also evaluated subjectively and MOS [STD] = 4.0 [0.19] (six people participated in the listening tests, STD represents standard deviation). The performance of the original IVA method is shown in Fig. 5*b*, it is clear from the results that the performance is poor because the CBSS algorithm cannot converge because of the limited number of samples truncate $(0.5Fs/T) = 3$ in each frequency bin.

In the second set of simulations, two tests are performed on the recorded mixtures of length = 7 s (for physically stationary sources case), which were separated by the original IVA method [54] and the RLSFIDI beamformer. The respective DOA (elevation and azimuth angles) obtained from 3D trackers are passed to the RLSFIDI beamformer and the resulting performance indices of the first test is shown in Fig. 6*a* and the performance of the

© The Institution of Engineering and Technology 2012

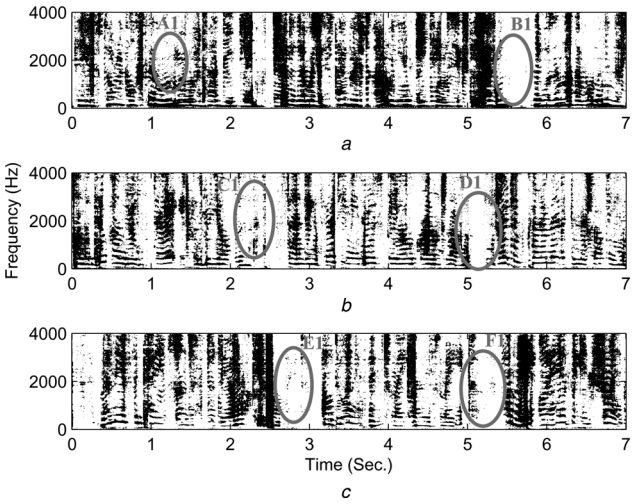


Fig. 7 *IBMs [35] of the three original speech signals used in the experiment at RT60 = 130 ms*

a Speaker 1
b Speaker 2
c Speaker 3
Highlighted areas, compared with the corresponding ones on Figs. 8 and 9 show how the post-filtering technique improves the output of the RLSFIDI beamformer

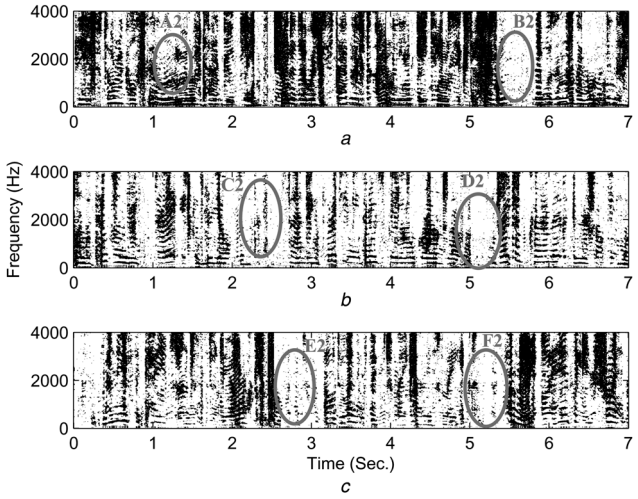


Fig. 8 *Binary masks of the speech signals separated by the RLSFIDI beamformer at RT60 = 130 ms*

a Speaker 1
b Speaker 2
c Speaker 3
Highlighted areas, compared with the corresponding original speech signals in Fig. 7 show that a considerable amount of interference from the other sources still exists when the ΔSINR = 14.97 dB

original IVA method for the same test is shown in Fig. 6*b*. The other objective evaluations for the both tests are shown in Table 2. These separations were also evaluated subjectively and MOS [STD]= 4.1 [0.15] and 4.2 [0.13]

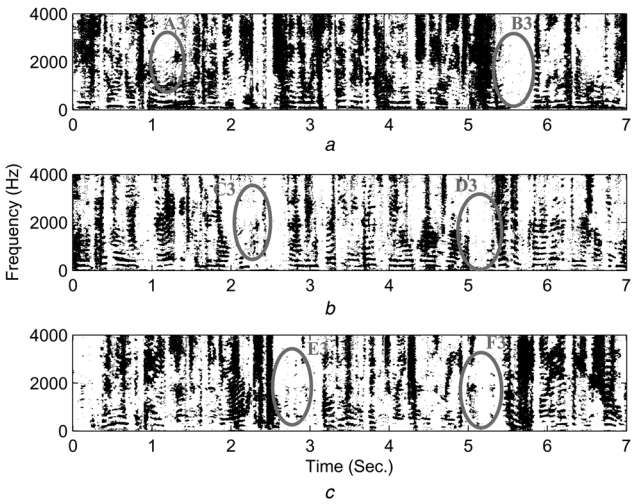


Fig. 9 *Binary masks of the three enhanced speech signals by the IBM T–F masking technique at RT60 = 130 ms*

a Speaker 1
b Speaker 2
c Speaker 3
Highlighted areas, compared with the corresponding ones on Figs. 7 and 8 show the post-filtering processing stage improves the output of the RLSFIDI beamformer. For these enhanced signals PEL = 10.15%, PNR = 11.22% and SINR = 16.83 dB

for the RLSFIDI beamformer and IVA methods, respectively. The performance of the higher-order statistics-based IVA method at RT60 = 130 ms with data length = 7 s is better than the RLSFIDI beamformer. The output of the RLSFIDI beamformer was further enhanced by the IBM technique. The masks of clean, estimated and enhanced speech signals are shown in Figs. 7–9, respectively. The highlighted areas, compared with the corresponding ones on Figs. 7–9 show how the post-filtering technique improves the speech signals separated by the RLSFIDI beamformer at the post-filtering process stage. In particular, the regions highlighted in Fig. 9 resemble closely the original sources in the regions shown in Fig. 7, the IBM technique has removed the granular noise shown in the regions highlighted in Fig. 8. The post-filtering enhanced the separated speech signals as shown in Table 3.

In the third set of simulations, two tests are performed on the generated mixtures of length = 7 s for RT60 = 300, 450 and 600 ms, which were separated by the RLSFIDI beamformer and the original IVA method [54]. The respective objective evaluations for each RT60 is shown in Table 4, which verifies the statement in [34] that at long impulse responses the separation performance of CBSS algorithms (based on second-order and higher-order statistics) is highly limited. For the condition $T > P$, we also increased the DFT length $T = 2048$ and there was no significant improvement observed because the number of samples in each frequency bin was reduced to truncate $(7Fs/T) = 27$.

Table 3 Objective evaluation: ΔSINR, SDR, SIR and SAR for the RLSFIDI beamformer after post-processing at RT60 = 130 ms, the length of the signals is 7 s[a]

	ΔSINR	SDR_1	SDR_2	SDR_3	SIR_1	SIR_2	SIR_3	SAR_1	SAR_2	SAR_3
RLSFIDI	16.83	6.59	7.90	8.04	14.30	14.37	15.54	7.55	9.17	9.02
beamformer	16.97	6.84	7.99	7.51	15.35	17.02	15.36	7.63	8.65	8.41

[a]Results are in decibels (dB)

© The Institution of Engineering and Technology 2012

doi: 10.1049/iet-spr.2011.0124

Table 4 Objective evaluation: ΔSINR, SDR, SIR and SAR for the RLSFIDI beamformer without post-processing and the original IVA method [54] for different RTs and when speakers are physically stationary for 7 s[a]

RT60, ms	Method	ΔSINR	SDR_1	SDR_2	SDR_3	SIR_1	SIR_2	SIR_3	SAR_1	SAR_2	SAR_3
300	RLSFIDI beamformer	11.25	2.93	0.90	6.91	9.31	4.87	8.97	4.54	4.35	11.66
		11.17	5.60	0.90	6.89	6.71	4.77	8.99	12.91	4.45	11.69
	IVA	12.02	0.96	3.49	5.44	1.74	5.08	7.28	10.99	9.70	10.81
		12.20	0.98	3.45	5.96	1.78	5.05	7.91	10.97	9.75	11.05
450	RLSFIDI beamformer	7.76	−0.77	0.45	4.26	4.94	3.94	7.75	1.78	4.51	7.52
		7.95	3.76	0.45	4.20	5.54	3.94	7.62	9.57	4.51	7.54
	IVA	6.55	−1.94	2.13	2.52	−0.37	4.26	4.70	6.44	7.62	7.75
		6.78	−0.86	2.98	4.09	0.43	5.31	6.33	7.81	7.90	8.93
600	RLSFIDI beamformer	6.30	−3.47	0.53	3.31	2.65	3.37	7.00	−0.37	3.9	4.85
		6.46	0.82	−0.12	2.24	4.19	3.37	6.87	4.89	3.96	4.88
	IVA	5.26	−5.26	−1.30	0.36	−2.88	3.99	5.77	3.17	1.67	2.80
		5.40	−5.14	−0.35	1.21	−3.44	2.80	4.04	4.82	4.35	4.89

[a]Results are in decibels (dB)

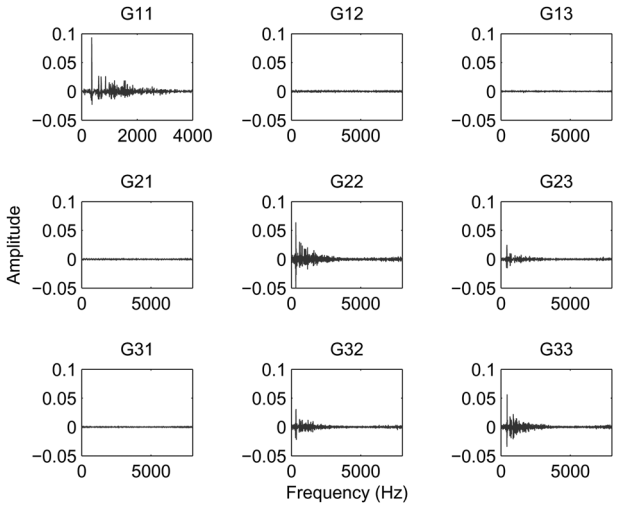


Fig. 10 *Combined impulse response G = WH by the original IVA method*

Reverberation time RT60 = 300 ms and SIR improvement was 12.2 dB

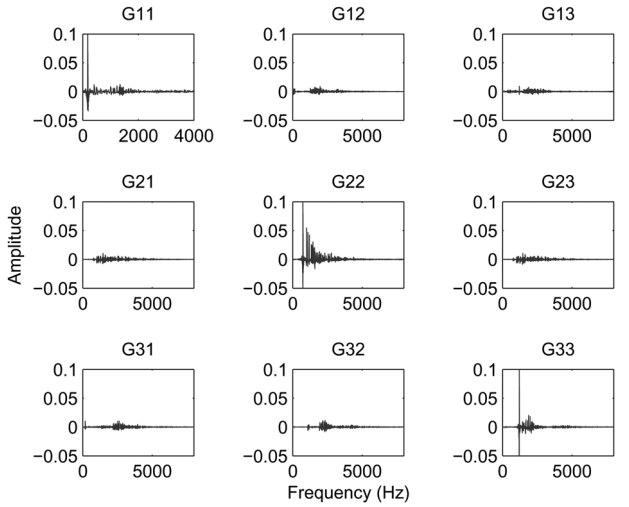


Fig. 11 *Combined impulse response G = WH by the RLSFIDI beamformer*

Reverberation time RT60 = 300 ms and SIR improvement was 11.2 dB

The justification of better performance for the RLSFIDI beamformer than the original IVA method, specially, at RT60 = 300 ms (Table 4) when Δ SINR of IVA method is higher than the RLSFIDI beamformer, is shown in Figs. 10 and 11. Actually, the CBSS method removed the interferences more effectively, therefore the ΔSINR is slightly higher. However, the separated speech signals are not perceptually so high quality, because the reverberations are not well suppressed. According to the 'law of the first wave front' [55], the precedence effect describes an auditory mechanism which is able to give greater perceptual weighting to the first wave front of the sound (the direct path) compared to later wave fronts arriving as reflections from surrounding surfaces. On the other hand, beamforming accepts the direct path and also suppresses the later reflections therefore the MOS is better. For comparison the typical room impulse response for RT60 = 300 ms is shown in Fig. 12.

In the final set of simulations, the separated speech signals by the RLSFIDI beamformer for each value of RT60 were further enhanced by applying the IBM technique. The respective objective evaluations for each RT60 are shown in Table 5. To show the performance of T–F masking as a post-processing stage, the results for RT60 = 300 ms for the first test are presented. The ideal binary masks (IBMs) of the three clean speech sources are shown in Fig. 13. In Fig. 14, the estimated binary masks (BM_s) of the output

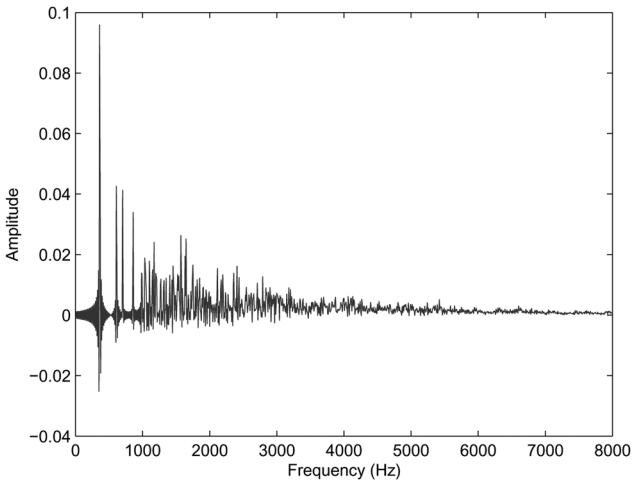


Fig. 12 *Typical room impulse response for reverberation time RT60 = 300 ms is provided for comparison*

© The Institution of Engineering and Technology 2012

Table 5 Final results: ΔSINR, SDR, SIR and SAR for the RLSFIDI beamformer after post-processing for different RTs and when speakers are physically stationary for 7 s[a]

RT60, ms	Method	ΔSINR	SDR_1	SDR_2	SDR_3	SIR_1	SIR_2	SIR_3	SAR_1	SAR_2	SAR_3
300	RLSFIDI beamformer	12.18	3.41	6.35	6.71	8.60	12.74	13.05	5.53	7.71	8.07
		12.36	5.22	5.80	6.50	13.00	11.47	16.54	5.05	7.48	7.05
450	RLSFIDI beamformer	8.86	−0.07	4.03	4.57	4.50	9.8	10.39	3.35	5.77	6.27
		9.76	4.86	3.85	4.09	10.19	9.62	15.32	2.88	5.64	4.56
600	RLSFIDI beamformer	7.59	−2.48	2.34	3.25	0.42	10.09	8.91	3.42	3.55	5.15
		7.91	−1.50	1.90	2.41	6.08	8.02	14.34	0.29	3.76	2.86

[a]Results are in decibels (dB)

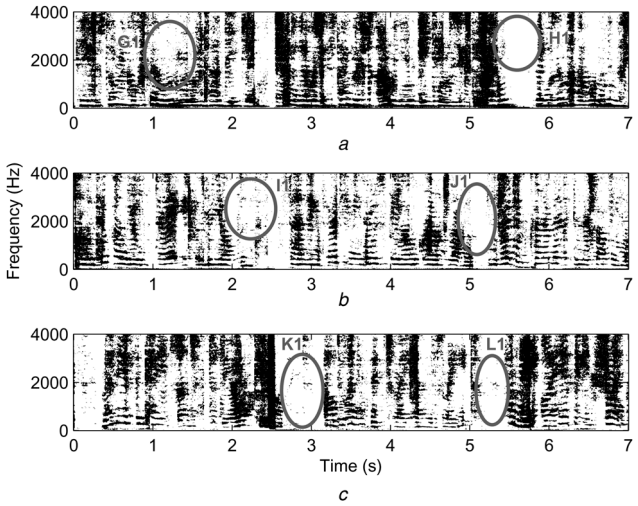


Fig. 13 *IBMs [35] of the three original speech signals used in the experiment at RT60 = 300 ms*

a Speaker 1
b Speaker 2
c Speaker 3
Highlighted areas, compared with the corresponding ones on Figs. 14 and 15 show how the post-filtering technique improves the output of the RLSFIDI beamformer

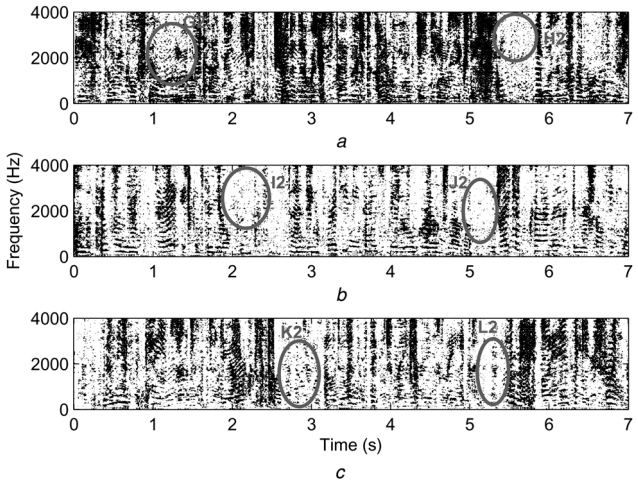


Fig. 14 *Binary masks of the speech signals separated by the RLSFIDI beamformer at RT60 = 300 ms*

a Speaker 1
b Speaker 2
c Speaker 3
Highlighted areas, compared with the corresponding original speech signals in Fig. 13 show that a considerable amount of interference from the other sources still exists when the ΔSINR = 11.25 dB

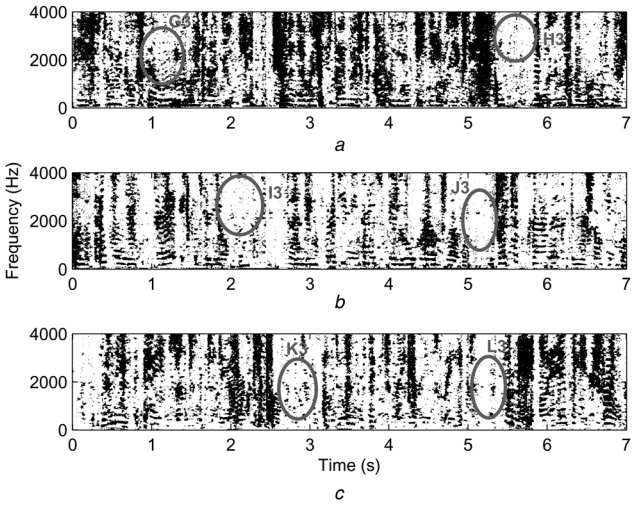


Fig. 15 *Binary masks of the three enhanced speech signals by the IBM T–F masking technique at RT60 = 300 ms*

a Speaker 1
b Speaker 2
c Speaker 3
Highlighted areas, compared with the corresponding ones on Figs. 13 and 14 show the post-filtering processing stage improves the output of the RLSFIDI beamformer. For these enhanced signals PEL = 24.82%, PNR = 28.04% and ΔSINR = 12.18 dB

signals obtained from the RLSFIDI beamformer are shown. These binary masks are applied on the spectrograms of the three selected microphones and masks of the enhanced speech signals are shown in Fig. 15. For the comparison, we show two regions in one of the three speech signals, which are marked as G_1, H_1, I_1, J_1, K_1, L_1 in the IBMs, G_2, H_2, I_2, J_2, K_2, L_2 in the SBM_s and G_3, H_3, I_3, J_3, K_3, L_3 in the final separated signals. From the highlighted regions, we can observe that the interference within one source that comes from the other is reduced gradually in the post-processing stage. The listening tests are also performed for each case and MOSs are presented in Table 6, which

Table 6 Subjective evaluation: MOS for the RLSFIDI beamformer with and without post-processing and the original IVA method, for different RTs, and when speakers are physically stationary for 7 s

RT60, ms	MOS [STD]		
	RLSFIDI beamformer	IVA method	proposed method
300	3.9 [0.16]	3.5 [0.17]	4.0 [0.21]
450	3.3 [0.19]	3.1 [0.15]	3.7 [0.15]
600	3.1 [0.20]	2.9 [0.31]	3.3 [0.15]

© The Institution of Engineering and Technology 2012

doi: 10.1049/iet-spr.2011.0124

indicates that at higher RT the performance of the RLSFIDI beamformer is better than the CBSS algorithms. The proposed solution not only improves the performance at lower RT, but also at higher RT60 when the performance of conventional CBSS algorithms is limited.

5 Conclusions

A novel multimodal (audio–visual) approach has been proposed for source separation of multiple physically moving and stationary sources in a reverberant environment. The visual modality was used to facilitate the source separation. The movement of the sources was detected with a 3D tracker based on an MCMC-PF, and the direction of arrival information of the sources to the microphone array was estimated. An RLSFIDI beamformer was implemented with linear and circular array configuration for a realistic 3D environment. The uncertainties in the source localisation and direction of arrival information were also controlled by using convex optimisation in the beamformer design. The proposed approach was shown to be a good solution to the separation of speech signals from multiple physically moving and stationary sources. For a highly reverberant environment, the performance of the RLSFIDI beamformer was enhanced by applying a binary T–F masking (IBM) technique in the post-filtering processing stage. The proposed approach has also been shown to provide a better separation than the conventional CBSS methods.

6 Acknowledgments

We thank the associate editor Professor Athanassios Manikas and the anonymous reviewers for their comments which have substantially improved the presentation of our work. We also thank the Engineering and Physical Science Research Council of the UK and the MOD University Defence Research Centre (UDRC) in Signal Processing for support.

This work was supported by the Engineering and Physical Science Research Council of the UK (grant number EP/H049665/1, EP/H050000/1, and EP/H012842/1), and the MOD University Defence Research Centre (UDRC) in Signal Processing.

7 References

1 Cherry, C.: 'Some experiments on the recognition of speech, with one and with two ears', *J. Acoust. Soc. Am.*, 1953, **25**, (5), pp. 975–979
2 Douglas, S.C., Sun, X.: 'Convolutive blind separation of speech mixtures using the natural gradient', *Speech Commun.*, 2003, **39**
3 Amari, S., Douglas, S.C., Cichocki, A., Yang, H.H.: 'Multichannel blind deconvolution and equalization using the natural gradient', *First IEEE Signal Processing Workshop on Signal Processing Advances in Wireless Communications*, 1997, pp. 101–104
4 Sawada, H., Mukai, R., Araki, S., Makino, S.: 'A robust and precise method for solving the permutation problem of frequency-domain blind source separation', *IEEE Trans. Speech Audio Process.*, 2004, **12**, (5), pp. 530–538
5 Ikram, M.Z., Morgan, D.R.: 'Permutation inconsistency in blind speech separation: Investigation and solutions', *IEEE Trans. Speech Audio Process.*, 2005, **13**, (1)
6 Hyvärinen, A., Karhunen, J., Oja, E.: 'Independent component analysis' (Wiley, New York, 2001)
7 Cichocki, A., Amari, S.: 'Adaptive blind signal and image processing: learning algorithms and applications' (John Wiley, 2002)
8 Wang, W., Sanei, S., Chambers, J.A.: 'Penalty function based joint diagonalization approach for convolutive blind separation of nonstationary sources', *IEEE Trans. Signal Process.*, 2005, **53**, (5), pp. 1654–1669
9 Parra, L., Spence, C.: 'Convolutive blind separation of non-stationary sources', *IEEE Trans. Speech Audio Process.*, 2000, **8**, (3), pp. 320–327
10 Makino, S., Sawada, H., Mukai, R., Araki, S.: 'Blind separation of convolved mixtures of speech in frequency domain', *IEICE Trans. Fundam.*, 2005, **E88-A**, (7), pp. 1640–1655
11 Nguyen, H.L., Jutten, C.: 'Blind source separation for convolutive mixtures', *Signal Process.*, 1995, **45**, pp. 209–229
12 Douglas, S.: 'Blind separation of acoustic signals (in microphone arrays: techniques and applications)' (Springer, Berlin, 2001)
13 Sumby, W., Pollack, I.: 'Visual contribution to speech intelligibility in noise', *J. Acoust. Soc. Am.*, 1954, **26**, pp. 212–215
14 McGurk, H., McDonald, J.: 'Hearing lips and seeing voices', *Nature*, 1976, **264**, pp. 746–748
15 Liew, A., Wang, S. (Eds.): 'Visual speech recognition: lip segmentation and mapping' (IGI, Global Press, 2009)
16 Cherry, C., Taylor, W.K.: 'Some further experiments upon the recognition of speech, with one and with two ears', *J. Acoust. Soc. Am.*, 1954, **26**, (4), pp. 554–559
17 Sanei, S., Naqvi, S.M., Chambers, J.A., Hicks, Y.: 'A geometrically constrained multimodal approach for convolutive blind source separation'. Proc. IEEE ICASSP, 2007, pp. 969–972
18 Asano, F., Yamamoto, K., Hara, I., *et al.*: 'Detection and separation of speech event using audio and video information fusion and its application to robust speech interface', *EURASIP J. Appl. Signal Process.*, 2004, **2004**, (11), pp. 1727–1738
19 Rivet, B., Girin, L., Jutten, C.: 'Mixing audiovisual speech processing and blind source separation for the extraction of speech signals from convolutive mixtures', *IEEE Trans. Audio Speech Lang. Process.*, 2007, **15**, (1), pp. 96–108
20 Naqvi, S.M., Yu, M., Chambers, J.A.: 'A multimodal approach to blind source separation of moving sources', *IEEE J. Sel. Top. Signal Process.*, 2010, **4**, (5), pp. 895–910
21 Wang, W., Cosker, D., Hicks, Y., Sanei, S., Chambers, J.A.: 'Video assisted speech source separation'. Proc. IEEE ICASSP, 2005, pp. 425–428
22 Naqvi, S.M., Zhang, Y., Tsalaile, T., Sanei, S., Chambers, J.A.: 'A multimodal approach for frequency domain independent component analysis with geometrically-based initialization'. Proc. EUSIPCO, Lausanne, Switzerland, 2008
23 Naqvi, S.M., Yu, M., Chambers, J.A.: 'A multimodal approach to blind source separation for moving sources based on robust beamforming'. Proc. IEEE ICASSP, Prauge, Czech Republic, 22–27 May 2011
24 Rivet, B., Girin, L., Jutten, C.: 'Visual voice activity detection as a help for speech source separation from convolutive mixtures', *Speech Commun.*, 2007, **49**, pp. 667–677
25 Tsalaile, T., Naqvi, S.M., Nazarpour, K., Sanei, S., Chambers, J.A.: 'Blind source extraction of heart sound signals from lung sound recordings exploiting periodicity of the heart sound'. Proc. IEEE ICASSP, Las Vegas, USA, 2008
26 Liu, Q., Wang, W., Jackson, P.: 'Use of bimodal coherence to resolve spectral indeterminacy in convolutive BSS'. Proc. LVA/ICA, St. Malo, France, 2010
27 Haykin, S., *et al.* (Eds.): 'New directions in statistical signal processing: from systems to brain' (The MIT Press, Cambridge, Massachusetts, London, 2007)
28 Mukai, R., Sawada, H., Araki, S., Makino, S.: 'Robust real-time blind source separation for moving speakers in a room'. Proc. IEEE ICASSP, Hong Kong, 2003
29 Koutras, A., Dermatas, E., Kokkinakis, G.: 'Blind source separation of moving speakers in real reverberant environment'. Proc. IEEE ICASSP, 2000, pp. 1133–1136
30 Prieto, R.E., Jinachitra, P.: 'Blind source separation for time-variant mixing systems using piecewise linear approximations'. Proc. IEEE ICASSP, 2005, pp. 301–304
31 Hild-II, K.E., Erdogmus, D., Principe, J.C.: 'Blind source extraction of time-varying, instantaneous mixtures using an on-line algorithm'. Proc. IEEE ICASSP, Orlando, Florida, USA, 2002
32 Anemuller, J., Gramss, T.: 'On-line blind separation of moving sound sources'. Proc. ICA, 1999
33 Addison, W., Roberts, S.: 'Blind source separtion with non-stationary mixing using wavelets'. Proc. Int. Conf. Independent Component Analysis and Blind Source Separation, 2006
34 Araki, S., Mukai, R., Makino, S., Nishikawa, T., Sawada, H.: 'The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech', *IEEE Trans. Speech Audio Process.*, 2003, **11**, (2), pp. 109–116
35 Pedersen, M.S., Liang, D., Larsen, J., Kjems, U.: 'Two-microphone separation of speech mixtures', *IEEE Trans. Neural Netw.*, 2008, **19**, (3), pp. 475–492

© The Institution of Engineering and Technology 2012

36 Jan, T., Wang, W., Wang, D.: 'A multistage approach to blind separation of convolutive speech mixtures', *Speech Commun.*, 2011, **53**, pp. 524–539

37 Wang, D.L., Brown, G.J.: 'Computational auditory scene analysis: principles, algorithms, and applications' (Wiley/IEEE Press, Hoboken, NJ, 2006)

38 Mabande, E., Schad, A., Kellermann, W.: 'Design of robust superdirective beamformers as a convex optimization problem'. Proc. IEEE ICASSP, Taipei, Taiwan, 2009

39 Khan, Z., Balch, T., Dellaert, F.: 'MCMC-based particle filtering for tracking a variable number of interacting targets', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2005, **27**, (11), pp. 1805–1818

40 Khan, Z., Balch, T., Dellaert, F.: 'MCMC data association and sparse factorization updating for real time multitarget tracking with merged and multiple measurements', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2006, **28**, (12), pp. 1960–1972

Q7 41 Viola, P., Jones, M.: 'Rapid object detection using a boosted cascade of simple features' 2001

42 Maganti, H.K., Gatica-Perez, D., McCowan, I.: 'Speech enhancement and recognition in meetings with an audio-visual sensor array', *IEEE Trans. Audio, Speech Lang. Process.*, 2007, **15**, (8), pp. 2257–2269

43 Ishii, Y., Hongo, H., Yamamoto, K., Niwa, Y.: 'Face and head detection for a real-time surveillance system'. Proc. 17th IEEE Conf. Pattern Recognition (ICPR), 2004, pp. 298–301

44 Tsai, R.Y.: 'A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf TV cameras and lenses', *IEEE J. Robot. Autom.*, 1987, **RA-3**, (4), pp. 323–344

45 Hartley, R., Zisserman, A.: 'Multiple view geometry in computer vision' (Cambridge University Press, 2001)

46 Van Veen, B., Buckley, K.: 'Beamforming: a versatile approach to spatial filtering', *IEEE ASSP Mag.*, 1988, **5**, (2), pp. 4–24

47 Parra, L.C.: 'Steerable frequency-invarient beamforming for arbitrary arrays', *J. Acoust. Soc. Am.*, 2006, **6**, pp. 3839–3847

48 Van Trees, H.L.: 'Detection, estimation, and modulation theory, Part IV, optimum array processing' (John Wiley and Sons, Inc., 2002)

49 Boyd, S., Vandenberghe, L.: 'Convex optimization' (Cambridge University Press, 2004)

50 Madhu, N., Breithaupt, C., Martin, R.: 'Temporal smoothing for spectral masks in the cepstral domain for speech separation'. Proc. IEEE ICASSP, 2008, pp. 45–48

51 Allen, J.A., Berkley, D.A.: 'Image method for efficently simulating small-room acousrics', *J. Acoust. Soc. Am.*, 1979, **65**, (4), pp. 943–950

52 Naqvi, S.M., Khan, M.S., Liu, Q., Wang, W., Chambers, J.A.: 'Multimodal blind source separation with a circular microphone array and robust beamforming'. Proc. European Signal Processing Conference (EUSIPCO'2011), Barcelona, Spain, 2011

53 Vincent, E., Fevotte, C., Gribonval, R.: 'Performance measuremet in blind audio source separation', *IEEE Trans. Speech Audio Process.*, 2006, **14**, pp. 1462–1469, http://sisec2010.wiki.irisa.fr/tiki-index.php

54 Kim, T., Attias, H., Lee, S., Lee, T.: 'Blind source separation exploiting higher-order frequency dependencies', *IEEE Trans. Audio Speech Lang. Process.*, 2007, **15**, (1)

55 Litovsky, R.Y., Colburn, H.S., Yost, W.A., Guzman, S.J.: 'The precedence effect', *J. Acoust. Soc. Am.*, 1999, **106**, pp. 1633–1654

© The Institution of Engineering and Technology 2012

doi: 10.1049/iet-spr.2011.0124

SPRSI20110124

Author Queries

S. Mohsen Naqvi, W. Wang, M. Salman Khan, M. Barnard, J.A. Chambers

Q1 Please compress the abstract within 200 words as per IEE style.

Q2 Please define DOA.

Q3 IEE style for matrices and vectors is to use bold italics. Please check that we have identified all instances.

Q4 Please check the change of Figs. 9 and 10 to Figs. 10 and 11 as in PDF file is ok.

Q5 Please provide page numbers in [2,5,54].

Q6 Please provide three authors' name with et al. if more than six authors are there, else provide all authors' name in [27].

Q7 Please provide the type (book/proceeding) of [41].