ACOUSTIC VECTOR SENSOR BASED REVERBERANT SPEECH SEPARATION WITH PROBABILISTIC TIME-FREQUENCY MASKING

Xionghu Zhong[⋆]*, Xiaoyi Chen*[‡†]*, Wenwu Wang*[†]*, Atiyeh Alinaghi*[†]*, and Annamalai B. Premkumar*[⋆]

[⋆] School of Computer Engineering, College of Engineering, Nanyang Technological University, Singapore, 639798.

[‡] Department of Acoustic Engineering, School of Marine Technology, Northwestern Polytechnical University, China, 710072.

[†] Centre for Vision, Speech and Signal Processing, Department of Electronic Engineering, University of Surrey, UK, GU2 7XH.

Emails: [⋆]{xhzhong and asannamalai}@ntu.edu.sg, [†‡]{w.wang, A.Alinaghi and xiaoyi.chen}@surrey.ac.uk.

ABSTRACT

Most existing speech source separation algorithms have been developed for separating sound mixtures acquired by using a conventional microphone array. In contrast, little attention has been paid to the problem of source separation using an acoustic vector sensor (AVS). We propose a new method for the separation of convolutive mixtures by incorporating the intensity vector of the acoustic field, obtained using spatially co-located microphones which carry the direction of arrival (DOA) information. The DOA cues from the intensity vector, together with the frequency bin-wise mixing vector cues, are then used to determine the probability of each time-frequency (T-F) point of the mixture being dominated by a specific source, based on the Gaussian mixture models (GMM), whose parameters are evaluated and refined iteratively using an expectation-maximization (EM) algorithm. Finally, the probability is used to derive the T-F masks for recovering the sources. The proposed method is evaluated in simulated reverberant environments in terms of signal-to-distortion ratio (SDR), giving an average improvement of approximately 1.5 dB as compared with a related T-F mask approach based on a conventional microphone setting.

Index Terms— Acoustic vector sensor, acoustic intensity, EM algorithm, blind source separation, direction of arrival.

1. INTRODUCTION

Speech source separation aims to estimate the desired speech signals in the presence of other speech signals or interfering sounds. It offers great potentials in many applications such as automatic speech recognition, teleconferencing and hearing aids. Traditionally, it is performed by using a microphone array together with estimation techniques developed based on the acoustic pressure measurements. Recently, a co-located sensor structure, namely acoustic vector sensor (AVS) is employed to measure the acoustic pressure as well as to calculate acoustic intensity [1], showing good performance on the estimation of direction of arrival (DOA) [2] and speech enhancement [3]. However, speech separation from sound mixtures acquired by an AVS has not been well studied, especially, under reverberant room environment.

In this paper, we propose a new method for separating reverberant speech mixtures by incorporating the intensity of the acoustic field that can be estimated from AVS recordings. Based on the acoustic intensity, we can extract the DOAs of the sources at each time-frequency (T-F) point of the mixture. We employ this DOA information together with the T-F bin-wise mixing vector cue to determine the probability of each T-F point of the mixture being dominated by a specific source, with the assumption that the sources are sparse in the T-F domain. The source occupation likelihood at each T-F point is evaluated from the mixtures based on the Gaussian mixture models (GMM), with the model parameters evaluated and refined iteratively by the expectation-maximization (EM) algorithm.

The main contribution lies in the use of the acoustic intensity information within the EM based probabilistic T-F masking technique for speech separation. Different from the EM method in [4], we have incorporated the acoustic intensity vector that allows the use of the bin-wise DOA cue. The remainder of this paper is organized as follows. In section 2, the AVS source separation model and the estimation of DOA and mixing vector cues based on the AVS model are introduced. Bin-wise T-F classification by combining these two cues using the EM algorithm is discussed in section 3. The experimental setup and results are given in section 4, followed by conclusions in section 5.

2. AVS BASED SOURCE SEPARATION IN THE REVERBERANT ENVIRONMENT

In this work, we assume that the sources and the sensor are strictly located at a 2-D ($x-y$) plane, i.e., the elevation angle of the sources are zero. Therefore, only two gradient components are included in a single AVS. As mentioned in [5], the true acoustic vector sensor should measure the pressure gradient directly, however, the gradient value can also be estimated indirectly by differentiating the measurements obtained from

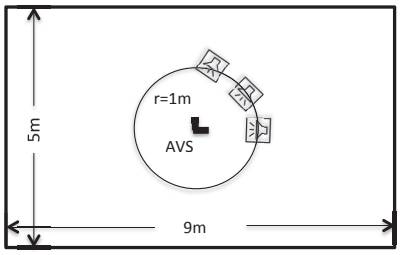
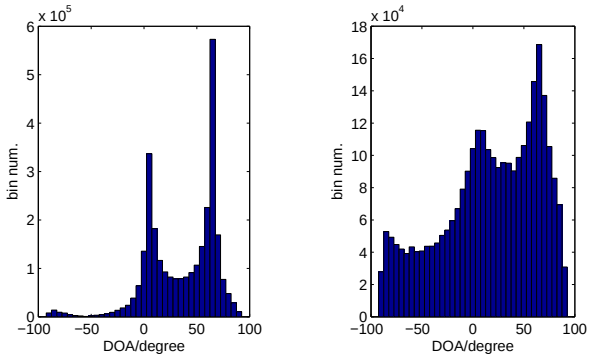


Fig. 1. An illustration of AVS and shoe-box room experiment environment.



(a) (b)

Fig. 2. The histogram of DOAs under (a) anechoic and (b) reverberant ($T_{60} = 0.35$s) environments. Two speech sources are simultaneously active at 5° and 65° respectively.

the pressure microphones, and the latter one is adopted in this paper. The geometry forming such an AVS in a shoe-box room is shown in Fig. 1. Three microphones are employed to construct an acoustic vector sensor for acoustic pressure measuring and pressure gradient calculation. In a noise-free room acoustic environment, the received mixtures from the source signals $s_n(t)$, for $n = 1, \ldots, N$ can be written as

$$\begin{bmatrix} p_0(t) \\ p_x(t) \\ p_y(t) \end{bmatrix} = \sum_{n=1}^{N} \begin{bmatrix} h_0^n(t) \\ h_x^n(t) \\ h_y^n(t) \end{bmatrix} \star s_n(t) \tag{1}$$

where N is the number of sources, t is the discrete time index, $\star$ denotes convolution, and $p_0(t)$, $p_x(t)$ and $p_y(t)$ are the acoustic pressure signal received from the sensors located at the origin, x-coordinate and y-coordinate respectively. In (1), $h_0^n(t)$, $h_x^n(t)$ and $h_y^n(t)$ represent the corresponding room impulsive response (RIR) from the nth source to the sensors cascading the direct path as well as the multipath responses.

The pressure gradient can then be obtained from the acoustic pressure as

$$\mathbf{g}(t) = \begin{bmatrix} g_x(t) \\ g_y(t) \end{bmatrix} = \begin{bmatrix} p_x(t) - p_0(t) \\ p_y(t) - p_0(t) \end{bmatrix} \tag{2}$$

where $g_x(t)$ and $g_y(t)$ is the pressure gradient corresponding to the x- and y- coordinates, respectively. The general form of the speech mixtures at the output of a single AVS can thus be constructed as $[p_0(t), \mathbf{g}(t)^T]^T$.

The aim of blind source separation (BSS) with the AVS settings is therefore to estimate source signals $s_n(t)$, $n = 1, \ldots, N$, from the mixtures $[p_0(t), \mathbf{g}(t)^T]^T$, without knowing RIRs, $h_0^n(t)$, $h_x^n(t)$ and $h_y^n(t)$ [1, 6]. To achieve this, we adopt the probabilistic T-F masking technique with the mask estimated using DOAs and mixing vector cues, as discussed in detail next.

2.1. Direction of arrival estimation with an AVS

In [5], Nehorai and Paldi assume that the signal behaves as a plane wave at the sensor. With this assumption, the acoustic particle velocity can be expressed as

$$\mathbf{v}(t) = -\frac{1}{\rho_0 c}\mathbf{g}(t) \odot \vec{\mathbf{u}} \tag{3}$$

where $\mathbf{v}(t) = [v_x(t), v_y(t)]^T$, $\odot$ denotes the element-wise product, ρ_0 is the ambient density of the air, c is the velocity of sound wave in the air, and $\vec{\mathbf{u}}$ is an unit vector denotes the direction in x- and y- coordinates, which holds an opposite direction of the DOA, i.e., $\vec{\mathbf{u}} = [\vec{u}_x, \vec{u}_y]^T$. The instantaneous intensity vector can be denoted as the product of the acoustic pressure and the particle velocity. By taking the short-time Fourier transform (STFT), the T-F representation of the intensity vector $\mathbf{I} = [I_x(\omega, k), I_y(\omega, k)]^T$ can be given as

$$I_x(\omega, k) = -\frac{1}{\rho_0 c}\left[\Re\{P_0^*(\omega, k)G_x(\omega, k)\}\vec{u}_x\right] \tag{4}$$

$$I_y(\omega, k) = -\frac{1}{\rho_0 c}\left[\Re\{P_0^*(\omega, k)G_y(\omega, k)\}\vec{u}_y\right] \tag{5}$$

where the superscript $*$ denotes conjugation, $\Re\{\cdot\}$ means taking the real part of its argument, ω and k are the frequency bin and time frame indices, and $P_0(\omega, k), G_x(\omega, k), G_y(\omega, k)$ are the STFTs of $p_0(t), g_x(t), g_y(t)$ respectively. The direction of the intensity can thus be obtained by

$$\theta(\omega, k) = \arctan\left[\frac{\Re\{P_0^*(\omega, k)G_y(\omega, k)\}}{\Re\{P_0^*(\omega, k)G_x(\omega, k)\}}\right] \tag{6}$$

Speech signal is, in general, sparse in the T-F domain [4], and as a result, it can be assumed that each T-F unit of the mixture is dominated by at most one source. The intensity direction $\theta(\omega, k)$ carries the DOA information of a source signal. By taking the histogram of $\theta(\omega, k)$ over all T-F points, the DOAs of sources, as shown in Fig. 2 (a), can be estimated and employed to separate the speech signals, to be explained in section 3.

2.2. Mixing vectors estimation

Different from the conventional microphone array, both the acoustic pressure and pressure gradient information are obtained at the output of the AVS. It was found experimentally in [6] that the performance will degrade when p_0 is used in source separation. Therefore, only the x- and y- gradient components of the AVS outputs are used to reconstruct the source signals. Assuming that only one source is dominant at each T-F unit, the STFT of the observations at the kth frame can be represented as

$$\begin{aligned}\mathbf{z}(\omega,k) &= \sum_{n=1}^{N}\hat{\mathbf{h}}_n(\omega)s_n(\omega,k)\\ &\approx \hat{\mathbf{h}}_n(\omega)s_n(\omega,k), \forall n\in[1,\ldots,N] \end{aligned} \quad (7)$$

where $\mathbf{z}(\omega,k)=[G_x(\omega,k),G_y(\omega,k)]^T$, $\hat{\mathbf{h}}_n(\omega)=[H_x^n(\omega)-H_0^n(\omega),H_y^n(\omega)-H_0^n(\omega)]^T$ and $H_0^n(\omega),H_x^n(\omega),H_y^n(\omega)$ are the STFTs of the $h_0^n(t),h_x^n(t),h_y^n(t)$ respectively, assuming a linear time-invariant (LTI) mixing system. Each observation vector is then normalized to remove the effect of the source amplitude. The mixing filter coefficients, $\hat{\mathbf{h}}_n$, are modeled by a complex Gaussian density (CGD) function, given as [7]

$$\begin{aligned}p(\mathbf{z}(\omega,k)|\hat{\mathbf{h}}_n(\omega),\gamma_n^2(\omega)) &= \frac{1}{\left(\pi\gamma_n^2(\omega)\right)^2}\\ \times\exp&\left(-\frac{||\mathbf{z}(\omega,k)-(\hat{\mathbf{h}}_n^H(\omega)\mathbf{z}(\omega,k))\hat{\mathbf{h}}_n(\omega)||^2}{\gamma_n^2(\omega)}\right)\end{aligned} \quad (8)$$

where $\hat{\mathbf{h}}_n$ is the centroid with a unit norm $||\hat{\mathbf{h}}_n(\omega)||^2=1$, and $\gamma_n^2(\omega)$ is the variance. The CGD function is evaluated for each observed T-F unit. The orthogonal projection of each observation $\mathbf{z}(\omega,k)$ onto the subspace spanned by $\hat{\mathbf{h}}_n$ can be estimated by $(\hat{\mathbf{h}}_n^H(\omega)\mathbf{z}(\omega,k))\hat{\mathbf{h}}_n(\omega)$. The minimum distance between the T-F unit $\mathbf{z}(\omega,k)$ and the subspace is thus $||\mathbf{z}(\omega,k)-(\hat{\mathbf{h}}_n^H(\omega)\mathbf{z}(\omega,k))\hat{\mathbf{h}}_n(\omega)||$ and represents the probability of that T-F point belonging to the nth source. The probability of each T-F unit coming from source n can thus be estimated to find out which source is dominant in that unit.

3. DOA AND MIXING VECTOR CUES BASED T-F ASSIGNMENT WITH EM ALGORITHM

With the increase of reverberations in the room environment, the DOA histogram will be blurred thus giving distorted direction information, as shown in Fig. 2 (b). To improve the reliability of allocating each T-F unit to a specific source, we propose to combine the DOA cue $\theta(\omega,k)$ with the T-F observations $\mathbf{z}(\omega,k)$. A GMM is then applied to the observation set. In GMM, a Gaussian distribution is employed for each source n, and thus *N* Gaussian distributions are mixed by the mixing weight $\psi_n(\omega)$. The main task is to find the model parameters (the mean and variances) that best fit the observations $\{\theta(\omega,k),\mathbf{z}(\omega,k)\}$. The parameter set Θ is given by

$$\Theta=\{\xi_n(\omega),\sigma_n^2(\omega),\hat{\mathbf{h}}_n(\omega),\gamma_n^2(\omega),\psi_n(\omega)\}$$

where ξ_n and σ_n^2 are the mean and variance of the DOAs, and $\hat{\mathbf{h}}_n(\omega)$ and $\gamma_n^2(\omega)$ are those of the mixing vector. Given an observation set, the parameters that maximize the log likelihood

$$\begin{aligned}L(\Theta)&=\max_{\Theta}\sum_{\omega,k}\log p(\theta(\omega,k),\mathbf{z}(\omega,k)|\Theta)\\ &=\max_{\Theta}\sum_{\omega,k}\log\sum_{n}[\psi_n(\omega)\mathcal{N}(\theta(\omega,k)|\xi_n(\omega),\sigma_n^2(\omega))\\ &\qquad\times\mathcal{N}(\mathbf{z}(\omega,k)|\hat{\mathbf{h}}_n(\omega),\gamma_n^2(\omega))]\end{aligned} \quad (9)$$

can be estimated using the EM algorithm [8] by iterating the E-step and the M-step until convergence .

In the E-step, given the estimated parameters, Θ at the M-step, and the observations, assuming the statistical independence [4], the probability that the nth source presents at each T-F unit of the mixture is calculated as

$$\begin{aligned}\nu_n(\omega,k)\propto\psi_n(\omega)&\mathcal{N}(\theta(\omega,k)|\xi_n(\omega),\sigma_n^2(\omega))\\ &\times\mathcal{N}(\mathbf{z}(\omega,k)|\hat{\mathbf{h}}_n(\omega),\gamma_n^2(\omega))\end{aligned} \quad (10)$$

where $\nu_n(\omega,k)$ is the occupation likelihood.

In the M-step, the DOA parameters $(\xi_n(\omega),\sigma_n^2(\omega))$ and the mixing vector parameters $(\hat{\mathbf{h}}_n(\omega),\gamma_n^2(\omega))$ are re-estimated for each source using the estimated occupation likelihood $\nu_n(\omega,k)$ in the E-step and the observations [4]. As there is usually no prior information about the mixing filters, for the first iteration, we set $\mathcal{N}(\mathbf{z}(\omega,k)|\hat{\mathbf{h}}_n(\omega),\gamma_n^2(\omega))=1$ in equation (10) to remove the effect of the mixing vector contribution. Once the mask $M_n(\omega,k)\equiv\nu_n(\omega,k)$ is obtained after one iteration based on only the information of DOA cue, the parameters of the mixing vectors, $(\hat{\mathbf{h}}_n(\omega),\gamma_n^2(\omega))$, can be estimated from the next M-step as follows

$$\mathbf{R}_n(\omega)=\sum_k\nu_n(\omega,k)\mathbf{z}(\omega,k)\mathbf{z}^H(\omega,k) \quad (11)$$

$$\gamma_n^2(\omega)=\frac{\sum_k\nu_n(\omega,k)||\mathbf{z}(\omega,k)-(\hat{\mathbf{h}}_n^H(\omega)\mathbf{z}(\omega,k))\hat{\mathbf{h}}_n(\omega)||^2}{\sum_k\nu_n(\omega,k)} \quad (12)$$

$$\psi_n(\omega)=\frac{1}{K}\sum_k\nu_n(\omega,k) \quad (13)$$

where K is the number of all time frames, and the optimum $\hat{\mathbf{h}}_n$ is the eigenvector corresponding to the maximum eigenvalue of $\mathbf{R}_n$.

After the convergence of the EM algorithm, the sources along x- and y- coordinates, i.e. $S_x^n(\omega,k)$ and $S_y^n(\omega,k)$, are finally recovered by using the masks $M_n(\omega,k)$ (i.e. the occupation likelihood described above) and the pressure gradient

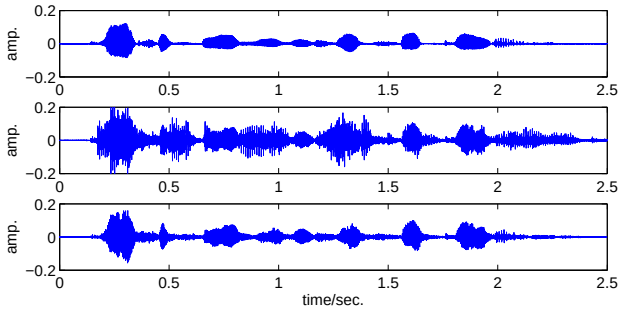


Fig. 3. Separation result under $T_{60} = 0.25$s: original signal (top), mixed signal (middle), and separated signal (bottom).

values at each coordinate

$$S_x^n(\omega, k) = M_n(\omega, k)G_x(\omega, k) \quad (14)$$

$$S_y^n(\omega, k) = M_n(\omega, k)G_y(\omega, k) \quad (15)$$

The time-domain speech sources are obtained by applying the overlapped inverse short-time Fourier transform (ISTFT) to S_x^n and S_y^n and then adding the x and y components of each source together.

It should be mentioned that the probabilistic classification in this BSS method is performed for each frequency bin separately and thus the permutation alignment over the frequency bins is still required. Rather than using a posteriori probability based approach as in [7], due to its high computational cost, we use the information from the DOA cue to solve the permutation alignment problem in the first iteration of the EM algorithm, similar to [4]. As a result, the remaining iterations of the EM algorithm will not be affected by the permutation problem.

4. EXPERIMENTS AND RESULTS

The proposed separation approach is tested for mixtures of two speech sources in simulated room environments. A shoebox room (as shown in Fig. 1) with a dimension of $9 \times 5 \times 3$ m^3 is employed. The separation performance is compared with that using the mixing vector cues obtained from the AVS and the conventional microphone array, respectively. In each experiment, the AVS and the microphone array are located at the center of the room. For AVS, the microphones at $x-$ and $y-$ coordinates are 0.5 cm away from the one at the origin. For the conventional microphone, we tested two different shapes for the array setup, respectively the 'L'-shaped array with the same spacing (0.5 cm from the origin) between the microphones as used in the AVS, and a uniform linear array (ULA) composed of three microphones. The former setup allows us to compare the performance of the proposed AVS based source separation method with the conventional microphone array based method with an identical microphone geometry and spacing. However, such a spacing (0.5 cm) is

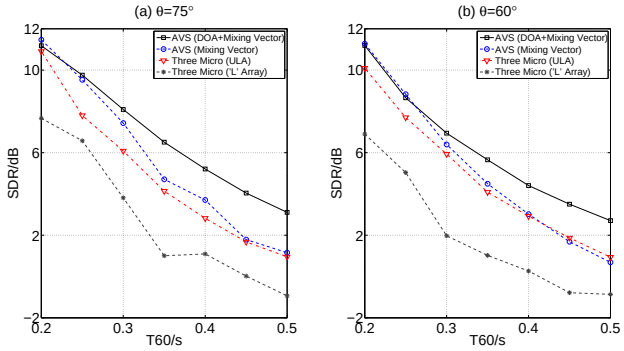


Fig. 4. SDR versus different T_{60}s and interference angles.

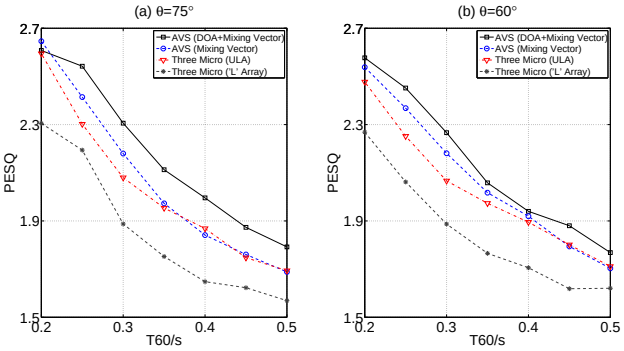


Fig. 5. PESQ versus different T_{60}s and interference angles.

rarely used in conventional microphone arrays due to its poor localisation performance (as shown in Fig. 4). For this reason, in our experiments, a larger spacing, i.e. neighbouring microphones spaced 5 cm apart, is used for the ULA, as done similarly in [6].

Similar to [4], 15 utterances are randomly chosen from the TIMIT dataset as source signals. These signals are shortened to 2.5 s for consistency. Moreover, all speech signals are normalized to have the same root mean square (RMS) amplitude before convolving with the RIRs, which are simulated by using the image method [9]. Different wall reflection coefficients are set to simulate different reverberant environments, which result in various T_{60}s from 200 ms to 500 ms with a step of 50 ms. To generate the mixtures, 15 pairs were chosen randomly from those 15 selected utterances. The target source was placed at 5° and the interferer at 50°, 65° and 80° respectively, leading to the angle difference between the target source and interferer θ at 45°, 60°, and 75°. Both the source and interference are located at 1 m from the microphones.

Fig. 3 gives an example of the separation result under $T_{60} = 250$ ms. It shows that the source signal can be recovered satisfactorily from the speech mixtures. The separation performance averaged over all the 15 mixtures is evaluated based on the signal-to-distortion ratio (SDR) [10] and perceptual evaluation of speech quality (PESQ) [11]. We applied

an FIR Wiener filter to the estimated signal with the target signal as reference. Therefore, any energy in the estimated signal that could be explained by a filtered version of the target signal was considered as the target signal. Any remaining energy was considered as distortion [4]. Fig. 4 shows SDR results of the proposed AVS method and the mixing vector cues based method [4] using the AVS and the conventional microphone array, for θ at 75° and 60° respectively. Results for θ at 45° are not shown here due to space constraint. Almost under all different T_{60}s and different interference angles, the proposed AVS method performs better than the mixing vector cues based methods, the SDR improvements are about 1 dB, 1.5 dB and 4 dB on average, compared with the mixing vector cues based source separation using the AVS, the ULA and the 'L'-shaped array, respectively. The corresponding PESQ improvements on average are 0.08, 0.15 and 0.3 respectively, as shown in Fig. 5.

It can be noticed from Fig. 4 and 5, the AVS based source separation method shows better performance than that using the conventional microphone array, however, the performance degrades rapidly without exploiting the DOA information, especially in the highly reverberant environments. In contrast, the proposed method which combines the DOA and mixing vector cues together shows more remarkable performance improvement for all the reverberation conditions tested.

5. CONCLUSION

We have presented a new algorithm for the separation of convolutive mixtures by incorporating the intensity vector of the acoustic field with probabilistic time-frequency masking. Instead of using a linear array, three microphones spatially co-located are employed to measure the acoustic intensity. The DOA cue and the mixing vector cue are then modeled by Gaussian mixture models for source separation. An EM algorithm is then introduced to estimate and refine the probability of each T-F point of the mixture belonging to each source. Simulation results in SDR and PESQ show the advantage of using AVS over the conventional microphone array for source separation, due to the high precision DOAs provided by the AVS. Future work includes applying the proposed approach to separate speech mixtures of more sources.

Acknowledgment

This work was supported by the Engineering and Physical Sciences Research Council, mainly under the grant EP/I000992/1, and in part under the grants EP/H012842/1 and EP/H050000/1.

6. REFERENCES

[1] B. Gunel, H. Hacihabiboglu, and A. M. Knondoz, "Acoustic source separation of convolutive mixtures based on intensity vector statistics," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 4, pp. 748–756, May 2008.

[2] X. Zhong and A. B. Premkumar, "Particle filtering approaches for multiple acoustic source detection and 2-D direction of arrival estimation using a single acoustic vector sensor," *IEEE Trans. Signal Process.*, vol. 60, no. 9, pp. 4719 –4733, Sept. 2012.

[3] M. Shujau, C. H. Ritz, and I. S. Burnett, "Speech enhancement via separation of sources from co-located microphone recordings," in *IEEE International conference on, Acoustics Speech and Signal Processing*. IEEE, 2010, pp. 137–140.

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

[5] A. Nehorai and E. Paldi, "Acoustic vector-sensor array processing," *IEEE Trans. Signal Processing*, vol. 42, no. 9, pp. 2481–2491, 1994.

[6] M. Shujau, C. H. Ritz, and I. S. Burnett, "Separation of speech sources using an acoustic vector sensor," in *IEEE International workshop on, Multimedia Signal Processing*, 2011, pp. 1–6.

[7] H. Sawada, S. Araki, and S. Makino, "A two-stage frequency-domain blind source separation method for underdetermined convolutive mixtures," *IEEE Workshop on Appl. Signal Process. Audio and Acoust.*, pp. 139–142, Oct. 2007.

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

[9] 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, Jul. 1979.

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

[11] L. D. Persia, D. Milone, H. Rufiner, and M.Yanagida, "Perceptual evaluation of blind source separation for robust speech recognition," *Signal Process.*, vol. 88, no. 10, pp. 2578–2583, Oct. 2008.