MULTI-SOURCE DIRECTION-OF-ARRIVAL ESTIMATION IN A REVERBERANT ENVIRONMENT USING SINGLE ACOUSTIC VECTOR SENSOR

Kai Wu, V. G. Reju, and Andy W. H. Khong

School of Electrical & Electronic Engineering,
Nanyang Technological University, Singapore,
Email: wu0001ai@e.ntu.edu.sg, {reju, andykhong}@ntu.edu.sg.

ABSTRACT

We address the problem of estimating direction-of-arrivals (DOAs) for multiple sound sources using a single acoustic vector sensor (AVS) in an enclosed room environment. It is well-known that multi-source DOA estimation in an enclosed environment is challenging due to room reverberation, environmental noise and overlapping of the source spectra. In this work, we propose a multi-source DOA estimation algorithm which exploits co-location of the sensor elements in AVS. We identify time-frequency (TF) zones of the received signals in which only one source is dominant with a high signal-to-reverberation ratio. DOA estimation is then achieved via the use of clustering of the Hermitian angle feature. Simulation results show that the proposed DOA estimation algorithm is robust to both reverberation and environmental noise.

***Index Terms*—** Multi-source DOA estimation, acoustic vector sensor, time-frequency sparsity, Hermitian angle

1. INTRODUCTION

Estimating the direction-of-arrivals (DOAs) of acoustic sources is important for automatic camera steering, beamforming, robotics and surveillance [1–3]. The presence of reverberation and background noise present challenges that need to be addressed in a realistic environment. In addition, estimation of DOAs of multiple and simultaneously active sources in an adverse environment is still an open problem. For such applications, conventional approaches often utilize an array of omni-directional microphones and DOA estimation is achieved by exploiting phase-delay information between the microphones [4]. However, conventional arrays often require a large aperture which presents limits in space-constrained applications.

An acoustic vector sensor (AVS) [5] which consists of one monopole pressure sensor element collocated with three orthogonally oriented dipole elements has drawn much interest in the research community. Unlike the conventional array which requires spacing between microphones, a single AVS can accomplish spatial filtering with a compact configuration. For DOA estimation using AVS, an initial work was presented in [5] in which the Cramér-Rao lower bound (CRLB) was derived for a free-space scenario with Gaussian additive noise. In addition, the intensity and velocity-covariance based DOA estimators were proposed using a single AVS. In [6], a maximum steered response power (SRP) estimator was proposed to generalize the intensity and velocity-covariance based algorithms. However, it was not clear that which aforementioned algorithm and its parameters attain the CRLB. In [7], a maximum likelihood estimator was proposed as a specific realization of the SRP algorithm and the optimal parameter which attains the CRLB can be obtained with knowledge of noise statistics. In [8], the reverberation effect was examined for the intensity-based estimator. The derived statistical model shows that the DOA estimate is biased in the presence of reverberation. In addition to the use of single AVS, an array of AVSs has also been employed to improve the DOA estimation performance by exploiting beamforming and subspace methods [9–13]. Quaternion based approach has been found to achieve a more accurate subspace decomposition for DOA estimation [14]. In [15–17], sound source tracking algorithms have been developed for AVS.

Although significant progress has been made, DOA estimation of multiple simultaneously active sources in a reverberant environment is still challenging. Conventional multiple signal classification (MUSIC) algorithm requires more number of sensors than the number of sources [18]. In addition, most of the aforementioned algorithms assume a free-space model [5–7, 9, 12, 13] and accuracy is expected to reduce with increasing reverberation time [8]. In recent studies, it is assumed that speech source signals are sparse in the time-frequency (TF) domain and hence DOA estimation can be achieved by clustering the single-source TF points [19–21]. While these methods can be directly extended to the AVS, the co-location structure of the AVS elements and its intrinsic advantages have not been fully exploited for multi-source DOA estimation in a reverberant environment.

In this work, a multi-source DOA estimation algorithm using a single AVS is proposed. Since the effect of reverberation varies with frequency bins, the proposed algorithm identifies low-reverberant-single-source (LRSS) zones in the TF domain of the received signals. These LRSS zones are defined as those with high signal-to-reverberation ratio and that only one source is dominant compared to the other sources. We proceed to explain why the proposed LRSS identification technique is well suited for the structure of AVS. After the identification of LRSS zones, we propose to exploit the Hermitian angle [22, 23] which partitions the identified LRSS zones into different clusters corresponding to different sources. Finally, DOA estimation is performed on each cluster.

2. MATHEMATICAL MODEL

Consider I active sound sources in a reverberant environment. Given an AVS with one monopole and three orthogonal dipole elements co-located at the origin, the received signals can be modeled as

$$\begin{bmatrix} x_{\mathrm{p}}(n) \\ \mathbf{x}_{\mathrm{v}}(n) \end{bmatrix} = \sum_{i=1}^{I} s_i(n) * \begin{bmatrix} h_{\mathrm{p},i}(n) \\ \mathbf{h}_{\mathrm{v},i}(n) \end{bmatrix} + \begin{bmatrix} e_{\mathrm{p}}(n) \\ \mathbf{e}_{\mathrm{v}}(n) \end{bmatrix}, \quad (1)$$

where $x_{\mathrm{p}}(n)$ and $\mathbf{x}_{\mathrm{v}}(n)$ are the omni-directional and the three orthogonal element outputs, respectively, n is the discrete time index,

$s_i(n)$ is the ith source signal, $h_{\mathrm{p},i}(n)$ is the impulse response from the ith source to the monopole pressure element, $\mathbf{h}_{\mathrm{v},i}(n)$ is a 3×1 impulse response sample vector from the ith source to the dipole elements and $*$ denotes convolution operator. The variables $e_{\mathrm{p}}(n)$ and $\mathbf{e}_{\mathrm{v}}(n)$ are defined as the noise signals. Using short-time Fourier transform (STFT), (1) can be represented as

$$\underline{\mathbf{x}}(k,m) = \sum_{i=1}^{I} \underline{s}_i(k,m)\underline{\mathbf{h}}_i(k) + \underline{\mathbf{e}}(k,m), \tag{2}$$

where $\underline{\mathbf{x}}(k,m) = [\underline{x}_{\mathrm{p}}(k,m),\ \underline{\mathbf{x}}_{\mathrm{v}}^T(k,m)]^T$ is the 4×1 STFT coefficient vector of the received signals, $\underline{s}_i(k,m)$ is the STFT coefficient of the ith source signal, $\underline{\mathbf{h}}_i(k) = [\underline{h}_{\mathrm{p},i}(k),\ \underline{\mathbf{h}}_{\mathrm{v},i}^T(k)]^T$ is the 4×1 vector formed by the STFT coefficients of the impulse responses, $\underline{\mathbf{e}}(k,m) = [\underline{e}_{\mathrm{p}}(k,m),\ \underline{\mathbf{e}}_{\mathrm{v}}^T(k,m)]^T$ is the vector of noise STFT coefficients, m is the frame index and k is the frequency-bin index.

In order to analyze the effect of reverberation, we further decompose $\underline{\mathbf{h}}_i(k)$ into direct-path component $\underline{\mathbf{h}}_i^{\mathrm{d}}(k)$ and reflection components $\underline{\mathbf{h}}_i^{\mathrm{r}}(k)$ such that (2) can be rewritten as

$$\underline{\mathbf{x}}(k,m) = \sum_{i=1}^{I} \underline{s}_i(k,m)\left[\underline{\mathbf{h}}_i^{\mathrm{d}}(k) + \underline{\mathbf{h}}_i^{\mathrm{r}}(k)\right] + \underline{\mathbf{e}}(k,m). \tag{3}$$

Since $\underline{\mathbf{h}}_i^{\mathrm{d}}(k)$ contains only component from the direction of the source and $\underline{\mathbf{h}}_i^{\mathrm{r}}(k)$ contains all reflected components that are dependent on the environment, they can be expressed as

$$\underline{\mathbf{h}}_i^{\mathrm{d}}(k) = e^{-\jmath\omega_k\tau_i}\mathbf{q}_i, \tag{4}$$

$$\underline{\mathbf{h}}_i^{\mathrm{r}}(k) = \sum_r \alpha_i^r e^{-\jmath\omega_k\tau_i^r}\mathbf{q}_i^r, \tag{5}$$

where $\mathbf{q}_i = [1,\ \mathbf{u}_i^T]^T$, $\mathbf{u}_i = [\cos\psi_i\cos\phi_i,\ \cos\psi_i\sin\phi_i,\ \sin\psi_i]^T$ define the sensor manifold pointing towards the ith source, ϕ_i and ψ_i are the azimuth and elevation direct-path incident angles, respectively. The variable τ_i denotes the direct-path time delay from the ith source to the sensor, $\omega_k = 2\pi k/K$ is the discrete angular frequency, $\mathbf{q}_i^r = [1,\ \mathbf{u}_i^{rT}]^T$, $\mathbf{u}_i^r = [\cos\psi_i^r\cos\phi_i^r,\ \cos\psi_i^r\sin\phi_i^r,\ \sin\psi_i^r]^T$ define the manifold pointing towards the rth reflection component, ϕ_i^r and ψ_i^r are the corresponding incident angles, τ_i^r is the time-delay of the reflection and α_i^r is the attenuation due to absorption at the room boundaries. The objective of this paper is therefore to estimate $\mathbf{u}_i$ which, in turn, provides DOA estimates of the sources.

3. THE PROPOSED ALGORITHM

It is well-known that the magnitude response of $\underline{\mathbf{h}}_i(k)$ varies across frequencies. This implies that the effect of reverberation varies across frequency bins. We therefore propose an algorithm to identify LRSS zones in the TF domain where only one of the source signals with high signal-to-reverberant ratio is dominant. Using our proposed approach for the identification of LRSS zones, we show that the co-location of the vector-sensor elements in AVS is well-suited for the detection of LRSS zones. The detected LRSS zones are then used for DOA estimation. The flow diagram of the proposed algorithm is illustrated in Fig. 1.

3.1. Detection of Low-reverberant-single-source zone

Consider a TF zone $\mathcal{Z}(k',m')$ of size $K_{\mathrm{z}} \times M_{\mathrm{z}}$ with zone index (k',m'), where K_{z} is the zone width across the frequency bins and M_{z} is the zone length across time frames. Assuming that the zone

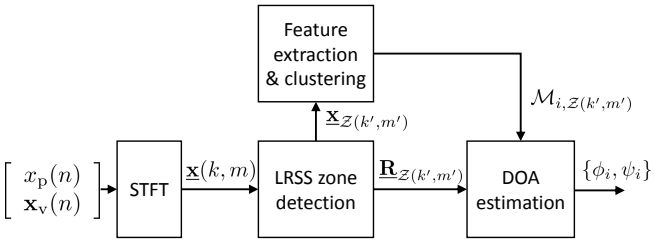

Fig. 1. Block diagram of the proposed DOA estimation algorithm.

has its centroid located at $\underline{\mathbf{x}}(k_{\mathrm{c}}, m_{\mathrm{c}})$ and that the zone has 50% overlap with adjacent zones, the zone indices are then given by $k' = \lfloor 2k_{\mathrm{c}}/K_{\mathrm{z}} \rceil$, $m' = \lfloor 2m_{\mathrm{c}}/M_{\mathrm{z}} \rceil$, where $\lfloor\cdot\rceil$ denotes the nearest integer. Within such a TF zone, if only the ith source is dominant and that the direct-path component is significantly larger than the reflection components and noise, the received signal in (3) can be approximated by

$$\underline{\mathbf{x}}(k,m) \approx \underline{s}_i(k,m)\underline{\mathbf{h}}_i^{\mathrm{d}}(k), \tag{6}$$

where $\underline{\mathbf{h}}_i^{\mathrm{d}}(k)$ is the direct-path component defined in (4). The covariance of $\underline{\mathbf{x}}(k,m)$ across all the TF points within the TF zone can then be estimated as

$$\begin{aligned}\underline{\mathbf{R}}_{\mathcal{Z}(k',m')} &= \mathbb{E}\{\underline{\mathbf{x}}(k,m)\underline{\mathbf{x}}^H(k,m)\} \\ &\approx \mathbb{E}\{|\underline{s}_i(k,m)|^2\underline{\mathbf{h}}_i^{\mathrm{d}}(k)\underline{\mathbf{h}}_i^{\mathrm{d}\,H}(k)\} \\ &= \sigma_i^2\mathbf{q}_i\mathbf{q}_i^T,\end{aligned} \tag{7}$$

where $\sigma_i^2 = \mathbb{E}\{|\underline{s}_i(k,m)|^2\}$ is the variance of the ith source signal and $\mathbb{E}\{\cdot\}$ denotes the expectation over the TF points within the TF zone $\mathcal{Z}(k',m')$.

It can be seen that for a LRSS zone, the rank of $\underline{\mathbf{R}}_{\mathcal{Z}(k',m')} \to 1$. On the contrary, as the number of sources or α_i^r in (5) increases, the rank of $\underline{\mathbf{R}}_{\mathcal{Z}(k',m')}$ will increase. Consider an example case in (3) where multiple sources are present in a reverberant-free environment. The covariance of $\sum_{i=1}^{I}\underline{s}_i(k,m)\underline{\mathbf{h}}_i^{\mathrm{d}}(k)$ will result in a rank that is greater than one as long as the sources are independent. In addition, in the presence of increased reverberation, the covariance of $\underline{s}_i(k,m)\left[\underline{\mathbf{h}}_i^{\mathrm{d}}(k) + \underline{\mathbf{h}}_i^{\mathrm{r}}(k)\right]$ will result a rank greater than one even if the TF zone corresponds to a single-source zone. This is due to the fact that $\underline{s}_i(k,m)\underline{\mathbf{h}}_i^{\mathrm{r}}(k)$ is linearly independent across frequency bins as described in (5) where the scaling factor $\alpha_i^r e^{-\jmath\omega_k\tau_i^r}$ for each $\mathbf{q}_i^r$ in the summation varies across frequencies.

Unlike conventional microphone arrays, it is important to note that the above rank-1 property of $\underline{\mathbf{R}}_{\mathcal{Z}(k',m')}$ for LRSS zone is derived from $\underline{\mathbf{h}}_i^{\mathrm{d}}(k)$ where the four channels of an AVS share the same phase delay. The direct-path component of the impulse response of a conventional microphone array is described by $\underline{\mathbf{h}}_i^{'\mathrm{d}}(k) = [e^{-\jmath\omega_k\tau_{i,1}},\ e^{-\jmath\omega_k\tau_{i,2}}, \cdots, e^{-\jmath\omega_k\tau_{i,P}}]^T$, where $\tau_{i,p}$ is the time-delay from the ith source to the pth microphone, and P is the number of microphones. It can be seen that $\underline{\mathbf{h}}_i^{'\mathrm{d}}(k)$ is linearly independent across frequency bins and therefore the corresponding $\underline{\mathbf{R}}_{\mathcal{Z}(k',m')}$ does *not* have the rank-1 property; detection of LRSS zones is thus not straightforward. It is also worth noting that the definition of $\underline{\mathbf{R}}_{\mathcal{Z}(k',m')}$ in (7) is different from [20]; equation (7) is defined as an average across time and frequencies, while the covariance in [20] is averaged across time frames only. It is indeed this manipulation of averaging across frequencies that exploits the common phase-delay property of the four channels in AVS for the detection of the LRSS zones.

To detect the LRSS zones, we divide the TF plane into zones of size $K_z \times M_z$ with 50% overlap between the zones across time frames and frequency bins. Each of these zones will be verified if they are LRSS zones by evaluating the rank of the corresponding covariance matrix. To determine whether the rank of the 4×4 covariance matrix $\underline{\mathbf{R}}_{\mathcal{Z}(k',m')}$ approaches to one, the coherence test [20]

$$\mathcal{C}_{\mathcal{Z}(k',m')} = \frac{1}{6} \sum_{a \neq b} \frac{|\underline{R}^{(a,b)}_{\mathcal{Z}(k',m')}|^2}{\underline{R}^{(a,a)}_{\mathcal{Z}(k',m')} \underline{R}^{(b,b)}_{\mathcal{Z}(k',m')}} \tag{8}$$

can be used, where $\mathcal{C}_{\mathcal{Z}(k',m')}$ is the coherence value for the zone $\mathcal{Z}(k', m')$, $\underline{R}^{(a,b)}_{\mathcal{Z}(k',m')}$ is the $(a,\ b)$ element of the matrix $\underline{\mathbf{R}}_{\mathcal{Z}(k',m')}$. In (8), $0 \leq \mathcal{C}_{\mathcal{Z}(k',m')} \leq 1$ and a higher value of $\mathcal{C}_{\mathcal{Z}(k',m')}$ implies that the rank of $\underline{\mathbf{R}}_{\mathcal{Z}(k',m')}$ is closer to 1. Hence, in order to detect the LRSS zones, we define a threshold $\mathcal{C}_{\text{thd}}$ such that zones with $\mathcal{C}_{\mathcal{Z}(k',m')} > \mathcal{C}_{\text{thd}}$ will be designated as LRSS zones. These zones will then be used for multi-source DOA estimation.

3.2. Feature extraction

Given a set of LRSS zones Γ that contains all of the identified LRSS zones $\{\mathcal{Z}(k', m')|\mathcal{Z}(k', m') \in \Gamma\}$ in a time block, these zones are clustered such that each cluster corresponds to a different source and DOA estimation can be performed on each cluster. To cluster the LRSS zones, the Hermitian angle will be exploited.

Theorem: The Hermitian angle between two arbitrary complex vectors $\mathbf{r}_1$ and $\mathbf{r}_2$ is defined as [22]

$$\theta = \cos^{-1}(|\cos(\theta_C)|), \tag{9}$$

where the cosine of complex-valued angle θ_C is given by $\cos(\theta_C) = \mathbf{r}_1^H \mathbf{r}_2 / \|\mathbf{r}_1\| \|\mathbf{r}_2\|$ and $\|\cdot\|$ denotes Euclidian norm. In addition, the Hermitian angle between $\mathbf{r}_1$ and $\mathbf{r}_2$ will remain the same even if the vectors are multiplied by any complex scalars [23].

To extract the feature corresponding to the source DOAs using Hermitian angle, we first take the TF point with the highest power in each LRSS zone, i.e.,

$$\underline{\mathbf{x}}_{\mathcal{Z}(k',m')} = \arg \max_{\underline{\mathbf{x}}(k,m) \in \mathcal{Z}(k',m')} ||\underline{\mathbf{x}}(k,m)||. \tag{10}$$

The elements of $\underline{\mathbf{x}}_{\mathcal{Z}(k',m')}$ are then used to form six two-element sub-vectors given by $\breve{\underline{\mathbf{x}}}^{<a,b>}_{\mathcal{Z}(k',m')} = [\underline{x}^{(a)}_{\mathcal{Z}(k',m')}, \underline{x}^{(b)}_{\mathcal{Z}(k',m')}]^T$, $\{a,\ b\} \subset \{\text{p},\ \text{v}_\text{x},\ \text{v}_\text{y},\ \text{v}_\text{z}\}$. Using (4), (6) can be rewritten as

$$\breve{\underline{\mathbf{x}}}^{<a,b>}_{\mathcal{Z}(k',m')} \approx \underline{s}_i(k,m) e^{-\jmath \omega_k \tau_i} \breve{\mathbf{q}}_i^{<a,b>}, \tag{11}$$

where $\breve{\mathbf{q}}_i^{<a,b>}$ is the vector consisting of the corresponding two elements of $\mathbf{q}_i$. In (11), $\breve{\underline{\mathbf{x}}}^{<a,b>}_{\mathcal{Z}(k',m')}$ is derived as a product of a source DOA dependent vector $\breve{\mathbf{q}}_i^{<a,b>}$ with a complex scalar. It is therefore expected that the Hermitian angle between $\breve{\underline{\mathbf{x}}}^{<a,b>}_{\mathcal{Z}(k',m')}$ and any reference vector $\mathbf{r}$ will be equal to the Hermitian angle between $\breve{\mathbf{q}}_i^{<a,b>}$ and $\mathbf{r}$. In other words, if the reference vector $\mathbf{r}$ is arbitrarily fixed, the Hermitian angles $\breve{\theta}^{<a,b>}_{\mathcal{Z}(k',m')}$ computed from $\breve{\underline{\mathbf{x}}}^{<a,b>}_{\mathcal{Z}(k',m')}$ will be uniquely determined by their dominant source DOAs.

Mathematically, the Hermitian angle between $\breve{\underline{\mathbf{x}}}^{<a,b>}_{\mathcal{Z}(k',m')}$ and an arbitrarily selected $\mathbf{r}$ can be computed, using (9), as

$$\breve{\theta}^{<a,b>}_{\mathcal{Z}(k',m')} = \cos^{-1}\left(\left| \frac{\mathbf{r}^H \breve{\underline{\mathbf{x}}}^{<a,b>}_{\mathcal{Z}(k',m')}}{\|\mathbf{r}\| \|\breve{\underline{\mathbf{x}}}^{<a,b>}_{\mathcal{Z}(k',m')}\|} \right| \right). \tag{12}$$

To improve clustering resolution, a 6×1 vector of Hermitian angles can be constructed by considering every pair of sensor elements, i.e.,

$$\mathbf{\Theta}_{\mathcal{Z}(k',m')} = \left[\breve{\theta}^{<\text{p},\text{v}_\text{x}>}_{\mathcal{Z}(k',m')},\ \breve{\theta}^{<\text{p},\text{v}_\text{y}>}_{\mathcal{Z}(k',m')},\ ...,\ \breve{\theta}^{<\text{v}_\text{x},\text{v}_\text{y}>}_{\mathcal{Z}(k',m')} \right]^T. \tag{13}$$

As discussed, $\mathbf{\Theta}_{\mathcal{Z}(k',m')}$ depends only on the DOAs of the dominant sources and thus can be used as a feature to cluster the LRSS zones.

3.3. Clustering and mask estimation

Given $\mathbf{\Theta}_{\mathcal{Z}(k',m')}$, the partitioning of $\mathbf{\Theta}_{\mathcal{Z}(k',m')}$ and hence the corresponding LRSS zones is performed in a multi-dimensional space. Any one of the well-established data clustering algorithms [24, 25], such as k-means [26] or fuzzy c-means (FCM) [27] may be used for this purpose. In this work, we employ the FCM algorithm which partitions the data into clusters with membership function that is inversely related to the distance of $\mathbf{\Theta}_{\mathcal{Z}(k',m')}$ to the centroid of each cluster. Therefore, defining i as the cluster index, the obtained membership function $\mathcal{M}_{i,\mathcal{Z}(k',m')}$ would be a smooth function.

For FCM, the number of clusters/sources must be known a priori. For an unknown number of clusters in practical scenarios, cluster validation techniques can be used [23, 28, 29]. These techniques assume that the maximum number of possible sources $I_{\max}$ is known. The variable $\mathbf{\Theta}_{\mathcal{Z}(k',m')}$ will then be clustered for $I = 2,\ \cdots,\ I_{\max}$, where I is the number of clusters. After each clustering, the cluster validity index will be computed and the I which achieves the optimal cluster validation index will be taken as the number of sources present. In this work, we assume that the number of sources is known since the focus of this work is DOA estimation.

3.4. DOA estimation

The membership function $\mathcal{M}_{i,\mathcal{Z}(k',m')}$ obtained from FCM algorithm is used as the mask for each source. Therefore, the covariance for the ith source is estimated by

$$\underline{\mathbf{R}}_i = \sum_{\mathcal{Z}(k',m') \in \Gamma} \mathcal{M}_{i,\mathcal{Z}(k',m')} \underline{\mathbf{R}}_{\mathcal{Z}(k',m')}. \tag{14}$$

In (14), the obtained $\underline{\mathbf{R}}_i$ is expected to contain only the ith source signal with low distortion since only the LRSS zones are taken into account. Since the TF points of the source signals are separated, any single-source DOA estimation algorithms, such as velocity-covariance based [5], maximum SRP estimator [6] or maximum likelihood estimator [7] can be applied. In this work we employ the MUSIC algorithm due to its high spatial resolution [18]. For a single source, the MUSIC spatial spectrum is defined as

$$\mathcal{J}_i(\mathbf{u}_\text{s}) = \frac{1}{\|\mathbf{q}_\text{s}^H \underline{\mathbf{U}}_i \underline{\mathbf{U}}_i^H \mathbf{q}_\text{s}\|}, \tag{15}$$

where $\mathbf{q}_\text{s} = [1,\ \mathbf{u}_\text{s}^T]^T$ with $\mathbf{u}_\text{s} = [\cos\psi\cos\phi,\ \cos\psi\sin\phi,\ \sin\psi]^T$ being the steering vector, and $\underline{\mathbf{U}}_i$ is the matrix consisting of three eigenvectors corresponding to the smallest eigenvalues of $\underline{\mathbf{R}}_i$. The direction of the ith source is then estimated by

$$\widehat{\mathbf{u}}_i = \arg \max_{\mathbf{u}_\text{s}} \mathcal{J}_i(\mathbf{u}_\text{s}),\ \text{ s.t. } \mathbf{u}_\text{s}^T \mathbf{u}_\text{s} = 1. \tag{16}$$

For different active sources, DOA estimation is performed using (14) to (16) for each of the identified clusters.

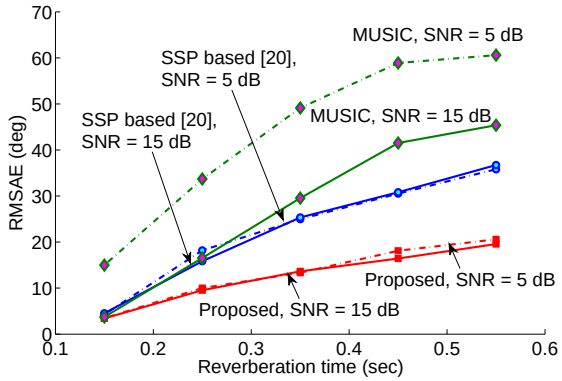

Fig. 2. RMSAE for different reverberation time and SNR, when two sources are present at $\phi_1 = 110^\circ$, $\psi_1 = -10^\circ$ and $\phi_2 = 165^\circ$, $\psi_2 = 15^\circ$.

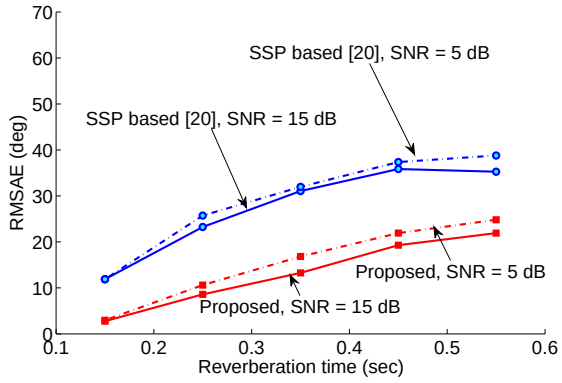

Fig. 3. RMSAE for different reverberation time and SNR, when three sources are present at $\phi_1 = 110^\circ$, $\psi_1 = -10^\circ$, $\phi_2 = 165^\circ$, $\psi_2 = 15^\circ$ and $\phi_3 = 220^\circ$, $\psi_3 = 20^\circ$.

4. SIMULATION RESULTS

Simulations were conducted for a $6\ \mathrm{m} \times 6\ \mathrm{m} \times 4\ \mathrm{m}$ room with an AVS located at $[3\ \mathrm{m},\ 3\ \mathrm{m},\ 1.3\ \mathrm{m}]$. Similar to [8], room impulse responses were generated using [30]. The pressure element was set as omni-directional and each of the vector-sensor elements was set as bi-directional with orthogonal orientation. Both male and female speech signals sampled at 16 kHz from the TIMIT database [31] were used as source signals. The sources were placed $1.7\ \mathrm{m}$ away from the sensor. White Gaussian noise at different signal-to-noise ratios (SNRs) were added to each of the four channels. The DOAs of the sources were estimated using 3 s block data in which the LRSS zones are identified. The frame length of STFT was 1024 samples. The TF-zone size was set to $62.5\ \mathrm{Hz} \times 256\ \mathrm{ms}$ and this corresponds to $K_z = 4$ and $M_z = 4$ with 50% overlap across frequency bins and time frames. The coherence test threshold was set to $\mathcal{C}_{\mathrm{thd}} = 0.75$ and the arbitrary selected reference vector for Hermitian angle computation was $\mathbf{r} = [1 + \jmath,\ 1 + \jmath]^T$.

In this work we compare the proposed algorithm with two existing multi-source DOA estimation algorithms. The conventional MUSIC algorithm [18] is used as baseline comparison in which the covariance matrix is computed without LRSS zone detection and clustering. The single-source point (SSP) based algorithm was also implemented by detecting and clustering the single-source points in TF plane [20]. It worth noting that the SSP based algorithm does not have the ability to detect the low-reverberant TF points/zones since the covariance matrix is obtained by averaging only across time frames (See Sec. 3.1 for explanation). The accuracy of DOA estimation is evaluated using angular error defined as the angle by which $\widehat{\mathbf{u}}$ deviates from $\mathbf{u}$ [5, 7]. For the case of multiple sources, it can be modified as $e = \frac{1}{I}\sum_{i=1}^{I} 2\sin^{-1}\left(||\widehat{\mathbf{u}}_i - \mathbf{u}_i||/2\right)$. We then quantify the performance across all the data blocks using the root-mean-square angular error (RMSAE) defined as $\mathrm{RMSAE} = \sqrt{\mathbb{E}\{e^2\}}$.

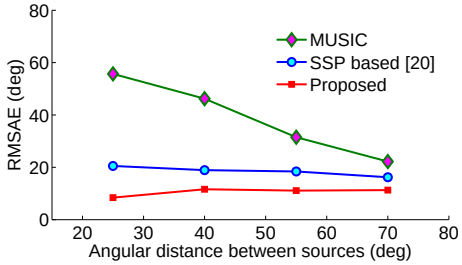

Fig. 4. Variation of RMSAE against angular distance between two active sources for $T_{60} = 300$ ms and SNR $= 15$ dB.

Figure 2 shows the variation of RMSAE against reverberation time for different noise levels when two sources are simultaneously active. The two sources are located at $\phi_1 = 110^\circ$, $\psi_1 = -10^\circ$ and $\phi_2 = 165^\circ$, $\psi_2 = 15^\circ$. These results show that the performance of the three algorithms degrade with increasing reverberation, as expected. While the MUSIC algorithm achieves an error of less than 5° when $T_{60} = 150$ ms and SNR $= 15$ dB, the performance deteriorates significantly with increasing reverberation and noise. The SSP based algorithm achieves lower error than the MUSIC algorithm since only single-source points are exploited for DOA estimation. The proposed algorithm achieves the lowest error compared to the other two algorithms. In addition, it is observed to be less sensitive to reverberation. This is due to the fact that for the proposed algorithm, only LRSS zones are identified which are less affected by reverberation.

Figure 3 shows the DOA estimation results for three active sources, where the third source is placed at $\phi_3 = 220^\circ$, $\psi_3 = 20^\circ$. In this figure, the results of MUSIC are not included since the MUSIC algorithm requires that the number of sources should be less than the number of dipole elements in AVS. Similar to previous simulation, the proposed algorithm achieves lower error than the SSP algorithm. However, the performance reduces with increasing reverberation since the number of LRSS zones is reduced.

Figure 4 illustrates the accuracy of the DOA estimation algorithms for various angular distance between two active sources. It can be observed that for the MUSIC algorithm, the error increases with reducing angular distance. Although MUSIC is well-known to achieve high resolution, its performance is degraded when only one AVS is used in a reverberant and noisy environment. On the other hand, the SSP based algorithm and the proposed algorithm are generally less sensitive to the source positions since they cluster the single-source TF points/zones before DOA estimation. The proposed algorithm achieves lower error than the SSP algorithm due to exploitation of LRSS zones.

5. CONCLUSION

We proposed a multi-source DOA estimation algorithm using a single AVS. The proposed algorithm identifies the LRSS zones available in the TF plane of the sensor outputs. The LRSS zones are then separated into clusters according to the sources for which the Hermitian angle feature is utilized. DOA estimation is then applied on each of these clusters. Simulation results show that the proposed algorithm achieves better performance than the MUSIC and SSP-based algorithms in a noisy and reverberant environment.

6. REFERENCES

[1] C. Zhang, D. Florêncio, D. E. Ba, and Z. Zhang, "Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings," *IEEE Trans. Multimedia*, vol. 10, no. 3, pp. 538–548, 2008.

[2] J. DiBiase, H. Silverman, and M. Brandstein, "Robust localization in reverberant rooms," *Microphone Arrays: Signal Processing Techniques and Applications.*, pp. 157–180, 2001.

[3] Y. Huang, J. Chen, and J. Benesty, "Immersive audio schemes," *IEEE Signal Process. Magazine*, vol. 28, pp. 20–32, Jan. 2011.

[4] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*, vol. 1, Springer, 2008.

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

[6] D. Levin, S. Gannot, and E. A. P. Habets, "Direction-of-arrival estimation using acoustic vector sensors in the presence of noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'11)*, 2011, pp. 105–108.

[7] D. Levin, E. A. P. Habets, and S. Gannot, "Maximum likelihood estimation of direction of arrival using an acoustic vector-sensor," *J. Acoust. Soc. Amer.*, vol. 131, no. 2, pp. 1240–1248, 2012.

[8] D. Levin, E. A. P. Habets, and S. Gannot, "On the angular error of intensity vector based direction of arrival estimation in reverberant sound fields," *J. Acoust. Soc. Amer.*, vol. 128, no. 4, pp. 1800–1811, 2010.

[9] M. Hawkes and A. Nehorai, "Acoustic vector-sensor beamforming and Capon direction estimation," *IEEE Trans. Signal Process.*, vol. 46, no. 9, pp. 2291–2304, 1998.

[10] M. Hawkes and A. Nehorai, "Wideband source localization using a distributed acoustic vector-sensor array," *IEEE Trans. Signal Process.*, vol. 51, no. 6, pp. 1479–1491, 2003.

[11] D. Rahamim, J. Tabrikian, and R. Shavit, "Source localization using vector sensor array in a multipath environment," *IEEE Trans. Signal Process.*, vol. 52, no. 11, pp. 3096–3103, 2004.

[12] H. Chen and J. Zhao, "Coherent signal-subspace processing of acoustic vector sensor array for DOA estimation of wideband sources," *Signal Processing*, vol. 85, no. 4, pp. 837–847, 2005.

[13] S. Zhao, S. Ahmed, Y. Liang, K. Rupnow, D. Chen, and D. L. Jones, "A real-time 3D sound localization system with miniature microphone array for virtual reality," in *Proc. 7th IEEE Int. Conf. Industrial Electronics and Applications (ICIEA)*, 2012, pp. 1853–1857.

[14] S. Miron, N. L. Bihan, and J. I. Mars, "Quaternion-MUSIC for vector-sensor array processing," *IEEE Trans. Signal Process.*, vol. 54, no. 4, pp. 1218–1229, 2006.

[15] 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, 2012.

[16] M. K. Awad and K. T. Wong, "Recursive least-squares source tracking using one acoustic vector sensor," *IEEE Trans. Aerospace and Electronic Systems*, vol. 48, no. 4, pp. 3073–3083, 2012.

[17] X. Zhong, A. B. Premkumar, and H. Wang, "Multiple wideband acoustic source tracking in 3-D space using a distributed acoustic vector sensor array," *IEEE Sensors Journal*, vol. 14, no. 8, pp. 2502–2513, Aug. 2014.

[18] D. H. Johnson and D. E. Dudgeon, *Array signal processing: concepts and techniques*, Simon & Schuster, 1992.

[19] W. Zhang and B. D. Rao, "A two microphone-based approach for source localization of multiple speech sources," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 8, pp. 1913–1928, 2010.

[20] S. Mohan, M. E. Lockwood, M. L. Kramer, and D. L. Jones, "Localization of multiple acoustic sources with small arrays using a coherence test," *J. Acoust. Soc. Amer.*, vol. 123, no. 4, pp. 2136–2147, 2008.

[21] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2193–2206, 2013.

[22] K. Scharnhorst, "Angles in complex vector spaces," *Acta Applicandae Math.*, vol. 69, no. 1, pp. 95–103, 2001.

[23] V. G. Reju, S. N. Koh, and I. Y. Soon, "Underdetermined convolutive blind source separation via time–frequency masking," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 1, pp. 101–116, 2010.

[24] A. K. Jain K, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Comput. Surveys*, vol. 31, no. 3, pp. 264–323, 1999.

[25] R. Xu and D. Wunsch II, "Survey of clustering algorithms," *IEEE. Trans. Neural Networks*, vol. 16, no. 3, pp. 645–678, 2005.

[26] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. the 5th Berkeley Symp. on Math. Statist. and Prob.*, 1967, vol. 1, pp. 281–297.

[27] J. C. Bezdek, *Pattern recognition with fuzzy objective function algorithms*, Kluwer Academic Publishers, 1981.

[28] H. Sun, S. Wang, and Q. Jiang, "FCM-based model selection algorithms for determining the number of clusters," *Pattern Recognition*, vol. 37, no. 10, pp. 2027–2037, 2004.

[29] Y. Zhang, W. Wang, X. Zhang, and Y. Li, "A cluster validity index for fuzzy clustering," *Information Sciences*, vol. 178, no. 4, pp. 1205–1218, 2008.

[30] E. A. P. Habets, "Room impulse response (RIR) generator," `http://home.tiscali.nl/ehabets/rir_generator.html`, (Accessed: 30/09/2014).

[31] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgrena, and V. Zue, *TIMIT Acoustic-Phonetic Continuous Speech Corpus*, Philadelphia, PA, 1993.