AN EXPECTATION-MAXIMIZATION EIGENVECTOR CLUSTERING APPROACH TO DIRECTION OF ARRIVAL ESTIMATION OF MULTIPLE SPEECH SOURCES

Xiong Xiao[1], Shengkui Zhao[2], Thi Ngoc Tho Nguyen[2], Douglas L. Jones[2], Eng Siong Chng[1,3], Haizhou Li[1,3,4]

[1]Temasek Lab@NTU, Nanyang Technological University, Singapore
[2]Advanced Digital Sciences Center, Singapore
[3]School of Computer Engineering, Nanyang Technological University, Singapore
[4]Department of Human Language Technology, Institute for Infocomm Research, Singapore
xiaoxiong@ntu.edu.sg, shengkui.zhao@adsc.com.sg, tho.nguyen@adsc.com.sg
dl-jones@illinois.edu, aseschng@ntu.edu.sg, hli@i2r.a-star.edu.sg

ABSTRACT

This paper presents an eigenvector clustering approach for estimating the direction of arrival (DOA) of multiple speech signals using a microphone array. Existing clustering approaches usually only use low frequencies to avoid spatial aliasing. In this study, we propose a probabilistic eigenvector clustering approach to use all frequencies. In our work, time-frequency (TF) bins dominated by only one source are first detected using a combination of noise-floor tracking, onset detection and coherence test. For each selected TF bin, the largest eigenvector of its spatial covariance matrix is extracted for clustering. A mixture density model is introduced to model the distribution of the eigenvectors, where each component distribution corresponds to one source and is parameterized by the source DOA. To use eigenvectors of all frequencies, the steering vectors of all frequencies of the sources are used in the distribution function. The DOAs of the sources can be estimated by maximizing the likelihood of the eigenvectors using an expectation-maximization (EM) algorithm. Simulation and experimental results show that the proposed approach significantly improves the root-mean-square error (RMSE) for DOA estimation of multiple speech sources compared to the MUSIC algorithm implemented on the single-source dominated TF bins and our previous clustering approach.

Index Terms— microphone arrays, direction of arrival, expectation-maximization, spatial covariance, eigenvector clustering.

1. INTRODUCTION

Direction of arrival (DOA) estimation of multiple speech sources using microphone arrays in noisy and reverberant environments has wide range of applications in areas such as distant automatic speech recognition [1, 2, 3, 4], teleconference systems [5], automatic camera steering [6], and hearing aids [7]. However, the performance of DOA estimation is greatly affected by the number of sources, room reverberation, background noises, and the configurations of microphone arrays. Reliable DOA estimation in such varying conditions for various microphone array setups is highly demanded.

When the number of sources is smaller than the number of sensors, the most widely used approaches for DOA estimation are subspace methods such as the multiple signal classification (MUSIC) [8], and its variants [9]. However, the performance of MUSIC drops significantly for large number of sources, high background noise level, or high reverberation level. When the number of sources is equal to or greater than the number of sensors, the DOA estimation has been addressed using the subspace approach [9, 10], the model-based approach [11, 12, 13], and the clustering approach [14, 15, 16, 17]. An example of model-based approach is the MESSL [12] for two-microphone array. In MESSL, probability distribution of interaural phase difference and interaural level difference are defined for each source and the estimation of the distribution parameters and assignment of time-frequency (TF) bins to sources can be solved by using an expectation-maximization (EM) algorithm iteratively. Although clustering methods [14, 16] can be applied to any array configurations, they require a strong assumption on the source signals that each TF bin is generated by only one source. Several methods have been used to select single-source dominant TF bins, such as the coherence test in [17] that selects TF bins with low-rank covariance matrices and the combination scheme in [15] which is found to be robust for multi-source DOA estimation in noisy and reverberant environments. A major limitation of the clustering methods is that they require a frequency normalization process to perform clustering across different frequencies. The normalization limits the frequency range used for the DOA estimation due to spatial aliasing. In order to use high frequencies, small aperture arrays have to be configured.

In this study, we propose an iterative eigenvector clustering approach for multi-source DOA estimation on the full frequency range without any constraint on the array configurations. The number of sources is assumed to be known. We first robustly identify single-source dominated TF bins based on TF bin selection scheme in [15]. We then obtain the largest eigenvectors of the covariance matrices corresponding to these TF bins. A mixture density function is used to model the eigenvectors and parameterized by source DOAs. To use eigenvectors in all frequencies, the steering vectors of sources in all frequencies are used in the model. As a result, the estimation of source DOAs becomes a density estimation problem and an EM algorithm is used to find the solution.

2. PROBLEM FORMULATION

Let us consider K far-field speech sources $s_k(t)$ $(k = 1, 2, ..., K)$ observed by an array of M microphones in a noisy and reverberant environment. The received signals can be modeled as:

$$x_m(t) = \sum_{k=1}^{K} \alpha s_k(t - \tau_{\theta_k,m}) + \sum_{k=1}^{K} s_k(t) \otimes h_m(t, \theta_k) + v_m(t), \quad (1)$$

where $m = 1, 2, ..., M$ is the microphone index; α is the propagation energy decay factor; $\tau_{\theta_k,m}$ is the time of arrival from the k^{th}

source location to the m^{th} microphone location; θ_k is the DOA of source $s_k(t)$; $\otimes$ denotes the convolution operation; $h_m(t, \theta_k)$ represents the impulse response of reverberation from source k to microphone m; $v_m(t)$ is the ambient noise.

In the short-time Fourier transform (STFT) domain, the model (1) can be rewritten as:

$$\begin{aligned} X_m(n,\omega) &= \sum_{k=1}^{K} \alpha S_k(n,\omega) e^{-j\omega\tau_{\theta_k,m}} \\ &+ \sum_{k=1}^{K} S_k(n,\omega) H_m(\omega,\theta_k) + V_m(n,\omega), \quad (2) \end{aligned}$$

where $n \in [1, N]$ is the time frame index; $\omega \in [0, \Omega - 1]$ is the frequency bin index; $X_m(n,\omega)$, $S_k(n,\omega)$, $V_m(n,\omega)$ are the frequency domain signals of $x_m(t)$, $s_k(t)$, and $v_m(t)$, respectively; $H_m(\omega,\theta_k)$ is the STFT transformed $h_m(t,\theta_k)$.

By stacking $X_m(n,\omega)$ for $m = 1, ..., M$, we can rewrite Equation (2) into a vector form as:

$$\mathbf{x}(n,\omega) = \sum_{k=1}^{K} S_k(n,\omega)[\mathbf{e}(\omega,\theta_k) + \mathbf{h}(\omega,\theta_k)] + \mathbf{v}(n,\omega), \quad (3)$$

where $\mathbf{e}(\omega,\theta_k) = [\alpha e^{-j\omega\tau_{\theta_k,1}}, ..., \alpha e^{-j\omega\tau_{\theta_k,M}}]^T$ is the steering vector of the array pointing to the DOA θ_k, and the normalized vector of $\mathbf{e}(\omega,\theta_k)$ by a reference channel is usually used. The vector $\mathbf{h}(\omega,\theta_k)$ is the reverberation vector from the DOA θ_k to the array. Our task is to estimate the DOAs of θ_k ($k = 1, 2, ..., K$) from the microphone signals (3).

3. PROPOSED EIGENVECTOR CLUSTERING APPROACH

In this section, we first describe the scheme for TF bin selection, and then the extraction of the eigenvectors of the covariance matrices of the selected TF bins. We then present a statistical framework for clustering the extracted eigenvectors. A generative model for the eigenvectors is introduced, and an expectation maximization (EM) algorithm is proposed to iteratively cluster the eigenvectors into a preseted number of sources.

3.1. Time-Frequency Bin Selection

When the signals are sufficiently sparse, the studies in [14, 16] assume there is one single source at each TF bin. For speech signals, this assumption is too strong as there can be many overlapped sources at some TF bins due to multiple speech sources and reverberation as shown in the signal model (3). Moreover, some TF bins may contain only noise. Thus, it is favorable to select the single source dominant TF bins for robust DOA estimation. In this study, we apply the TF-bin selection scheme described as follows. First, a noise-floor tracking algorithm is used to eliminate noise-only TF bins; then an onset detection algorithm is used to detect the direct-path TF bins; after that the coherence test in [17] is applied to the sample covariance matrix to identify low rank TF bins. The selected TF bins from this scheme are considered as single-source dominant TF bins. The detail of this scheme is presented in [15].

3.2. Eigenvector Extraction

The study in [16] clusters directly a frequency-normalized version of the TF bin $\mathbf{x}(n,\omega)$, which cannot eliminate the effects of the noise components embedded in the TF bins. We consider only the single-source dominant TF bins, $\mathbf{x}(n,\omega)$, selected by the above scheme, and compute the sample covariance matrix of this TF bin from $2C+1$ adjacent time-blocks as follows:

$$\tilde{\mathbf{R}}(n,\omega) = \frac{1}{2C+1} \sum_{c=n-C}^{n+C} \mathbf{x}(c,\omega)\mathbf{x}^H(c,\omega). \quad (4)$$

As there is only one dominant source s_k ($k \in 1, 2, ..., K$) which is from the direct path and above noise level, the above covariance matrix can be approximated as:

$$\tilde{\mathbf{R}}(n,\omega) \approx \sigma_k^2(n,\omega)\mathbf{e}(\omega,\theta_k)\mathbf{e}^H(\omega,\theta_k) + \sigma_V^2(n,\omega)\mathbf{I}, \quad (5)$$

where $\sigma_k^2(n,\omega)$ and $\sigma_V^2(n,\omega)$ denote the sample averaged power spectra of s_k and ambient noise, respectively. The noise is assumed to be identically and independently distributed. There is no reverberation term in (5) as only direct path TF bins are selected.

Taking eigendecomposition of $\tilde{\mathbf{R}}(n,\omega)$, the largest eigenvector $\mathbf{q}(n,\omega)$ of $\tilde{\mathbf{R}}(n,\omega)$ is obtained, which is known as the signal space and is highly correlated with the steering vector $\mathbf{e}(\omega,\theta_k)$. We propose to cluster these eigenvectors instead of the frequency-normalized phase of the eigenvectors in [15]. The advantage of the eigenvector clustering is that it allows the full use of the signal frequency range and weighted contribution of each eigenvector. Next, we present an EM-based eigenvector clustering approach.

3.3. Generative Model

The task is to cluster the eigenvectors of the selected TF bins into K clusters and find the DOAs of the clusters. In the following, we use $\mathbf{q}_{n\omega}$ as a short-form of $\mathbf{q}(n,\omega)$ for ease of presentation. Based on the graphical model representation, a generative model of the eigenvectors of TF bins is illustrated in Fig. 1. For each selected TF bin, we introduce a latent variable $\mathbf{z}_{n\omega}$ that specifies which source generates the TF bin. $\mathbf{z}_{n\omega}$ is a K dimensional vector where only the k^{th} element $z_{n\omega k}$ is 1 and all other elements are 0 if the TF bin is generated by the k^{th} source. The generative process works as follows. For the eigenvector located at time n and frequency ω, we first sample $\mathbf{z}_{n\omega}$ according to the discrete probability distribution function:

$$\begin{aligned} p(z_{n\omega k} = 1) &= \pi_k, \quad k \in [1, K], \\ \sum_{k=1}^{K} \pi_k &= 1. \quad (6) \end{aligned}$$

where π_k is the prior probability of source k. Suppose the sampled value is $z_{n\omega j} = 1$ which specifies the j^{th} source generates the TF bin, we then sample $\mathbf{q}_{n\omega}$ from the probability distribution associated with the j^{th} source which is defined as:

$$p(w_{n\omega}, \mathbf{q}_{n\omega} | z_{n\omega j} = 1; \theta_j) = \frac{\exp(\beta w_{n\omega} |\mathbf{q}_{n\omega}^H \mathbf{e}_{j\omega}|)}{\mathcal{E}(\beta,\theta_j)} \quad (7)$$

where $|\cdot|$ denotes the magnitude of a complex number and $\cdot^H$ denotes Hermitian transpose. $\mathbf{e}_{j\omega}$ is the $M \times 1$ steering vector of the source j at frequency ω and a function of both frequency and the source DOA θ_j. As $\mathbf{q}_{n\omega}$ is estimated from distorted signals in practice, we also introduce a positive scalar variable $w_{n\omega}$ that describes the reliability of the eigenvector. In our implementation, $w_{n\omega}$ is the ratio between the largest eigenvalue and the average of the rest eigenvalues. The denominator $\mathcal{E}(\beta,\theta_j)$ is defined as

$$\mathcal{E}(\beta,\theta_j) = \int_{\mathbf{q}_{n\omega}, w_{n\omega}, \omega} \exp(\beta w_{n\omega} |\mathbf{q}_{n\omega}^H \mathbf{e}_{j\omega}|) d\mathbf{q}_{n\omega} dw_{n\omega} d\omega \quad (8)$$

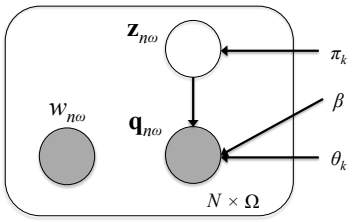

Fig. 1. Graphical model for the proposed eigenvector clustering

The exponential function in (7) makes $p(w_{n\omega}, \mathbf{q}_{n\omega}|z_{n\omega j} = 1; \theta_j) \geq 0$ for all possible value of $\mathbf{q}_{n\omega}$ and $w_{n\omega}$, and the use of denominator $\mathcal{E}(\beta, \theta_j)$ makes the distribution function integrates to 1. Hence, $p(w_{n\omega}, \mathbf{q}_{n\omega}|z_{n\omega j} = 1; \theta_j)$ is a valid probability density function (PDF). β is a scalar variable to control the spread of the PDF.

The PDF in (7) relies on the term $|\mathbf{q}_{n\omega}^H \mathbf{e}_{j\omega}|$ which reaches its maximum value of 1 if $\mathbf{q}_{n\omega} = \mu \mathbf{e}_{j\omega}$ with a scaling factor μ, which means that the eigenvector perfectly matches the source steering vector. The term also equals to its minimum value of 0 if the two vectors are orthogonal. Hence, the term $|\mathbf{q}_{n\omega}^H \mathbf{e}_{j\omega}|$ is a good measure of the similarity between the eigenvector and the source j. A higher similarity will produce a higher probability of the eigenvector being generated by the source.

The integration in (8) integrates over all possible value of $w_{n\omega}$, $\mathbf{q}_{n\omega}$, and ω, so it is intractable in practice. We make an assumption that $\mathcal{E}(\beta, \theta_j) = \mathcal{E}(\beta)$, so it does not depend on the source DOAs. It will be shown that it is not necessary to compute $\mathcal{E}(\beta)$ in the eigenvector clustering method, which significantly simplifies the complexity of the clustering method.

Given (6) and (7), we can write the PDF of the observed eigenvectors and reliablity variables as a mixture density function:

$$\begin{aligned} p(w_{n\omega}, \mathbf{q}_{n\omega}; \Theta) &= \sum_{k=1}^{K} p(z_{n\omega k} = 1) p(w_{n\omega}, \mathbf{q}_{n\omega}|z_{n\omega k} = 1; \theta_k) \\ &= \sum_{k=1}^{K} \pi_k \frac{\exp(\beta w_{n\omega} |\mathbf{q}_{n\omega}^H \mathbf{e}_{k\omega}|)}{\mathcal{E}(\beta)}. \end{aligned} \quad (9)$$

where Θ is the set of all source DOAs.

3.4. EM Algorithm for Eigenvector Clustering

The source DOAs can be estimated by maximizing the likelihood of the observed data in (9), which include both the eigenvectors and their reliability variables. Due to the latent variables $\mathbf{z}_{n\omega}$ which are not directly observed, there is no closed form solution to the problem. Hence, we propose an EM-based iterative algorithm. In the expectation (E) step, we compute the posterior probabilities of source k for each eigenvector observation using the existing estimated DOAs. By the Bayesian theorem, we have

$$p(z_{n\omega k} = 1|w_{n\omega}, \mathbf{q}_{n\omega}; \Theta') = \frac{\pi_k \exp(\beta w_{n\omega} |\mathbf{q}_{n\omega}^H \mathbf{e}_{k\omega}|)}{\sum_{j=1}^{K} \pi_j \exp(\beta w_{n\omega} |\mathbf{q}_{n\omega}^H \mathbf{e}_{j\omega}|)} \quad (10)$$

where Θ' is the set of source DOAs estimated from the previous iteration. In the first iteration of the EM algorithm, Θ' are randomly sampled in the range [0,359].

At the maximization (M) step, we find a new estimation of the source DOAs by maximizing the EM auxiliary function

$$\hat{\theta}_k = \arg \max_{\theta_k \in [0,359]} Q(\theta_k; \Theta'), \quad k \in [1, K] \quad (11)$$

$$\begin{aligned} Q(\theta_k; \Theta') &= \sum_{\{n,\omega\} \in \Psi} \gamma_{n\omega k} \log p(w_{n\omega}, \mathbf{q}_{n\omega}|z_{n\omega k} = 1; \theta_k) \\ &= \sum_{\{n,\omega\} \in \Psi} \gamma_{n\omega k} \log \frac{\exp(\beta w_{n\omega} |\mathbf{q}_{n\omega}^H \mathbf{e}_{k\omega}|)}{\mathcal{E}(\beta)}. \end{aligned} \quad (12)$$

where Ψ is the set that contains the time-frequency pairs of the selected TF bins and $\gamma_{n\omega k} = p(z_{n\omega k} = 1|w_{n\omega}, \mathbf{q}_{n\omega}; \Theta')$ is the posterior probability computed in the E step. We only allow the source DOAs to take integer values as real applications usually do not require higher resolution. As $\mathcal{E}(\beta)$ does not depend on source DOA, we can remove it form the optimization problem:

$$\hat{\theta}_k = \arg \max_{\theta_k \in [0,359]} \sum_{\{n,\omega\} \in \Psi} \gamma_{n\omega k} \beta w_{n\omega} |\mathbf{q}_{n\omega}^H \mathbf{e}_{k\omega}|) \quad (13)$$

The optimal DOA can be found by a grid search, i.e. computing the term in (13) for all the 360 possible values of θ_k and set $\hat{\theta}_k$ to the one producing the highest auxiliary function value.

In practical implementation, we reduce the computational complexity of the proposed clustering algorithm by setting $\gamma_{n\omega k}$ to 1 for the source of the highest probability and 0 for the rest. This reduces by K times the computation in (13), which is the most expensive part of the clustering algorithm. Note that β has no effect in the simplified algorithm.

3.5. Discussion on Frequency Range and Spatial Aliasing

One advantage of the proposed method is that it is able to use eigenvectors from TF bins of all frequencies. This is due to the fact that the steering vector and the eigenvector of the same frequency are used in computing the probability in (7). For each source cluster, the steering vectors of all frequencies for the cluster source DOA collectively act as the template of the cluster. This is different from other clustering methods such as k-means where only one template is used per cluster.

The posterior probability $p(z_{n\omega k} = 1|w_{n\omega}, \mathbf{q}_{n\omega}; \Theta')$ will have multiple peaks at high frequencies due to spatial aliasing. Hence, there is ambiguity in determining the source DOA if we only rely on high frequencies. However, this has limited effect on the proposed DOA estimation method as the DOA of a source is estimated from TF bins in all frequencies that "belong" to the source. Despite the spatial aliasing, the higher frequency TF bins still contain information that is useful for DOA estimation.

4. EXPERIMENTS

The proposed eigenvector clustering approach is evaluated and compared to the MUSIC subspace approach [8] and the clustering approach [15] for 2-dimensional DOA estimation on both simulated and real data. An 8-channel circular array with a diameter of 20cm is used for the tests. For all the testing data, the sampling rate is 16 kHz; the FFT length is 512, and the overlap size is 256. The performance of the tested approaches is evaluated using the RMSE.

For the simulated data, the image method [18] is used to generate the room impulse responses (RIRs) from specified source positions to the array positions. The array outputs are synthesized by convolving clean speech signals from the WSJCAM0 corpus [19] with the

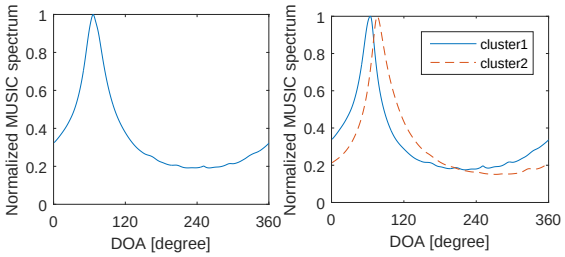

Fig. 2. Illustration of the normalized MUSIC spatial spectra computed from all selected TF bins (left figure) and clustered TF bins. The reverberation time is 1.0s and SNR=19.8dB.

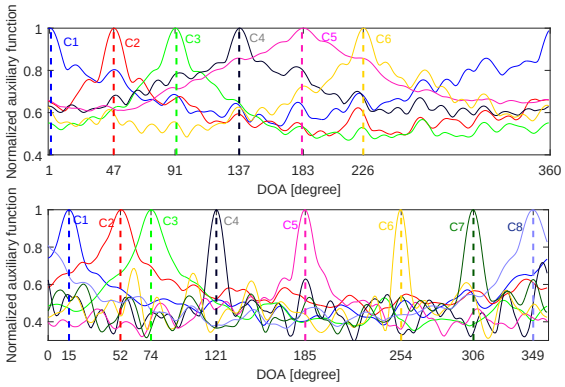

Fig. 3. Illustration of the normalized values of auxiliary function using the proposed eigenvector clustering approach for six real sources (pantry room, T60=0.47s, above figure) and for eight simulated sources (T60=1s, SNR=20dB, below figure).

generated RIRs. The additive noises from the REVERB Challenge 2014 corpus [20] are added to the array outputs. Different array outputs are randomly mixed to create the scenarios of multiple speech sources from different DOAs where the DOAs are separated by at least 5°. The audio file lengths are from 3s to 5s long. The reverberation times are 0.3s for small room, 0.6s for medium room, and 1.0s for large room. The distances between the sources and the array are 1.5m for small room, and 2m for medium and large rooms.

For the real data records, 8 GRAS40 PH microphones are configured for the real 8-channel circular array. NI cDAQ-9139 is used for audio capturing and analog-to-digital conversion (ADC). The recording is made in three different environments: small meeting room (4m×3m×2.5m), pantry room (6m×5m×2.5m), and lift lobby (8m×4m×3m). The measured reverberation times of the three environments are 0.34s, 0.47s, and 1.07s, respectively. We use one male speaker and one female speaker for the recording. The DOAs recorded for the male speaker are 0°, 45° and 90° at each environment, and the DOAs recorded for the female speaker are 135°, 180° and 225° at each environment. Different array outputs are mixed to create the scenarios of multiple speech sources. The audio files are cut to 6s long for each testing case.

4.1. Results

To illustrate the effectiveness of the proposed clustering approach, we plot the normalized MUSIC spatial spectra of two close sources ($\theta_1 = 65°$, $\theta_2 = 72°$) in Fig. 2. When the MUSIC spatial spectrum is computed from all selected TF bins, there is only one peak, therefore MUSIC fails to estimate the two directions. The MUSIC spatial spectrum computed from clustered TF bins obtained by the proposed clustering method shows two peaks located near the true source directions. This demonstrates the ability of the proposed method in resolving sources with close DOAs.

Fig. 3 illustrates the obtained normalized values of auxiliary function (13) for 6 real sources at $\{\theta_1, ..., \theta_6\} = \{2°, 47°, 92°, 137°, 182°, 226°\}$ and 8 simulated sources at $\{\theta_1, ..., \theta_8\} = \{13°, 50°, 74°, 118°, 186°, 254°, 306°, 349°\}$. The peaks of the auxiliary functions match well the source DOAs.

Table 1 shows the RMSE results for the simulated data with two mixed speech sources. It is observed that the proposed method outperforms other methods in all test conditions. All the RMSEs of the proposed approach are less than 10° while the RMSEs of MUSIC and the algorithm [15] are greater than 10° in general.

Table 2 shows the RMSE results of the real data with two mixed speech sources. The DOA ground truth was obtained using MUSIC for each source separately. It is observed that the proposed approach achieves much smaller RMSEs than MUSIC and the algorithm [15]. The results confirm that the clustering on full frequency range has significant advantage over the clustering on low-frequency range where the algorithm [15] uses 2.24kHz. It is seen that the RMSEs for the data of the small room are greater than the RMSEs for the data of the pantry room. It might be because there were strong reflections from the tables close to the microphone array during data collection in the small room. All algorithms produce higher RMSEs for the more reverberant lift lobby environment.

Table 1. The RMSE (in degrees) results for the simulated data.

Room	Method	SNR=20dB	SNR=10dB	SNR=0dB
Small	MUSIC	15.51	18.30	23.76
	Algorithm[14]	14.26	18.41	20.50
	Proposed	**2.78**	**2.84**	**6.75**
Medium	MUSIC	13.67	9.41	12.37
	Algorithm[14]	18.19	10.97	14.12
	Proposed	**3.76**	**5.15**	**3.07**
Large	MUSIC	7.87	4.99	14.94
	Algorithm[14]	11.70	10.10	11.49
	Proposed	**1.17**	**2.04**	**9.47**

Table 2. The RMSE (in degrees) results for the real data.

		Testing Environment		
		small	pantry	lift
Method	MUSIC	33.43	28.7	62.66
	Algorithm[14]	40.74	42.92	70.02
	Proposed	**1.7**	**0.92**	**13.02**

5. CONCLUSIONS

We have presented an eigenvector clustering approach for DOA estimation of multiple speech sources based on a probabilistic model and EM algorithm. The proposed model allows the full use of signal frequencies despite spatial aliasing in high frequencies. The approach can be applied to different sizes and configurations of microphone arrays. Experimental results using a 8-channel circular array show significantly reduced RMSEs of the proposed approach compared to the MUSIC algorithm and the clustering approach that uses only lower frequency range. The proposed approach is also shown to be capable for estimating DOAs of large number of speech sources.

6. REFERENCES

[1] M. Woelfel and J. McDonough, *Distant Speech Recognition*, Wiley, 2009.

[2] Steve Renals, Thomas Hain, and Hervé Bourlard, "Recognition and understanding of meetings the ami and amida projects," in *IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU*, Kyoto, 12 2007, IDIAP-RR 07-46.

[3] Emmanuel Vincent, Jon Barker, Shinji Watanabe, Jonathan Le Roux, Francesco Nesta, and Marco Matassoni, "The second CHiME speech separation and recognition challenge: Datasets, tasks and baselines," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 126–130.

[4] X. Xiao, S. Zhao, D. H. H. Nguyen, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "The NTU-ADSC systems for reverberation challenge 2014," in *Proceedings of the Reverberation Challenge Workshop*, 2014.

[5] 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 *Proceedings of the 7th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, 2012.

[6] S. Zhao, E. S. Chng, N. T. Hieu, and H. Li, "A robust real-time sound source localization system for olivia robot," in *Proceeding of the APSIPA Annual Summit and Conference*, 2010.

[7] B. Widrow, "A microphone array for hearing aids," in *Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium.*, 2000, pp. 7–11.

[8] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propagat.*, vol. AP-34, pp. 279–280, March 1986.

[9] S. Zhao, T. Saluev, and D. L. Jones, "Underdetermined direction of arrival estimation using acoustic vector sensor," *Signal Processing*, vol. 100, pp. 160–168, 2014.

[10] W.-K. Ma, T.-H. Hsieh, and C.-Y. Chi, "DOA estimation of quasi-stationary signals with less sensors than sources and unknown spatial covariance: a Khatri-Rao subspace approach," *IEEE Transactions on Signal Processing*, vol. 58, pp. 2168–2180, April 2010.

[11] Michael I. Mandel, Daniel P. W. Ellis, and Tony Jebara, "An EM algorithm for localizing multiple sound sources in reverberant environments," in *Advances in Neural Information Processing Systems*, B. Schölkopf, J. Platt, and T. Hoffman, Eds., pp. 953–960. MIT Press, Cambridge, MA, 2007.

[12] Michael Mandel, Ron J Weiss, Daniel PW Ellis, et al., "Model-based expectation-maximization source separation and localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 382–394, 2010.

[13] C. Liu, BC. Wheeler, WD. Jr. O'Brien, RC. Bilger, CR. Lansing, and AS. Feng, "Localization of multiple sound sources with two microphones," *Journal of Acoustic Society of America*, vol. 108(4), pp. 1888–1905, Oct 2000.

[14] S. Richard and O. Rilmaz, "On the approximation w-disjoint orthogonality of speech," in *Proceedings of 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2002, vol. 1.

[15] TNT Nguyen, S. Zhao, and Douglas L. Jones, "Robust DOA estimation of multiple speech sources," in *Proceedings of 2014 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014, pp. 2287–2291.

[16] S. Araki, H. Sawada, R. Mukai, and S. Makino, "DOA estimation for multiple sparse sources with normalized observation vector clustering," in *Proceedings of 2006 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2006, vol. 5.

[17] S. Mohan, M. E. Lockwood, M. L. Kramer, and D. L. Jones, "Localization of multiple acoustic sources with small arrays using a coherence test," *Journal of Acoustic Society of America*, vol. 123(4), pp. 2136–2147, April 2008.

[18] J.B. Allen and D.A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943Ű–950, April 1979.

[19] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJ-CAM0: a british english speech corpus for large vocabulary continuous speech recognition," in *Proceeding of ICASSP*, 1995, pp. 81–84.

[20] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas, S. Gannot, and B. Raj, "The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA-13)*, 2013.