ROBUST DOA ESTIMATION OF MULTIPLE SPEECH SOURCES

Nguyen Thi Ngoc Tho[⋆] *Shengkui Zhao*[⋆] *Douglas L. Jones*[⋆†] [*]

[⋆] Advanced Digital Science Center (ADSC),
Illinois at Singapore, 138632, Singapore
[†] Department of Electrical and Computer Engineering,
University of Illinois at Urbana-Champaign, IL 61820, United States

ABSTRACT

It is challenging to determine the directions of arrival of speech signals when there are fewer sensors than sources, particularly in noisy and reverberant environments. The coherence test by Mohan *et al.* exploits the time-frequency sparseness of non-stationary speech signals to select more relevant time-frequency bins to estimate directions of arrival. With no prior knowledge about the incoming sources, this work proposes a combination of noise-floor tracking, onset detection and a coherence test to robustly identify time-frequency bins where only one source is dominant. After that, the largest eigenvectors of covariance matrices corresponding to these bins are clustered and the directions of arrival of the sources are estimated based on the cluster centroids. Simulation and experimental results show that this method is able to localize 8 sources with small errors using only 3 omnidirectional microphones. The proposed method is robust to background noise and reverberation.

Index Terms— coherence test, direction of arrival estimation, eigenvector, microphone array, time-frequency

1. INTRODUCTION

Direction-of-arrival (DOA) estimation of acoustic sources has a wide range of applications in surveillance, tracking, teleconferencing and hearing aids. While DOA estimation is a mature field, challenging unsolved problems remain the subject of on-going research. One of those problems is to determine DOA when the number of sources exceeds the number of sensors, i.e., the under-determined DOA estimation.

Speech signals are known to be non-stationary and sparse in the time-frequency (TF) domain. Various research has exploited these two properties to solve the under-determined DOA estimation of speech sources. The majority of these researches further simplify the problem by assuming that "a single frequency bin is occupied by only a single source". Using this assumption, Rickard and Yilmaz [1] estimate the DOA at each TF bin based on a two-dimensional histogram of relative amplitude and delay parameters across time and frequency. With the same assumption, Araki *et al.* [2] cluster normalized TF bins and estimate the DOAs from the cluster centroids. Zhang *et al.* [3] assume frequency-modulated sources and estimate DOAs by averaging covariance matrices from TF bins of the same source. In order to solve for problems where the prior knowledge of the source is unknown, Mohan *et al.* [4] use a coherence test to select the rank-1 TF bins containing only one dominant source. With this coherence test, the aforementioned assumption is no longer required.

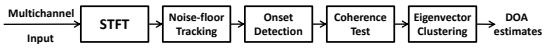

Fig. 1. Block diagram of the DOA estimation algorithm.

Another problem of DOA estimation is reverberation. While humans are able to localize sources in heavily reverberant environments, current DOA estimation techniques can only deal with a moderate amount of reverberation. Wallach *et al.* [5] show that humans have such ability because of the precedence effect: when the lags between the first arriving sound and subsequent sounds are sufficiently short (< 40 ms), all the sounds are fused into a single sound and the perceived direction of arrival is determined solely by the first arriving sound (onset). Huang *et al.* [6] apply the precedence effect and detect onsets to estimate DOAs of 2 sources using 3 microphones. Another problem that needs to be addressed is the presence of noise. Since speech signals are sparse in the time-frequency domain, some TF bins contain only noise; thus, it is favorable to eliminate these TF bins using noise-floor tracking to improve the accuracy of DOA estimation.

This work aims to mitigate the effect of noise and reverberation in the under-determined DOA problem and reduce the computation time. The key point is to incorporate noise-floor tracking and onset detection along with a coherence test to select the most relevant TF bins. By clustering the largest eigenvectors of the corresponding covariance matrices, the DOAs can be directly estimated from the cluster centroids. This method demonstrates robust performances for both simulation and experimental data.

[*]This study is supported by a research grant for the Human Sixth Sense Programme at the Advanced Digital Sciences Center from Singapore's Agency for Science, Technology and Research (A*STAR).

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

2. SIGNAL MODEL

All the sources are assumed to be independent and stationary over small intervals of time ($<$ 20 ms). These assumptions are approximately valid for speech signals. The system is assumed to be linear and time invariant so that the observed signal at the microphone array, $\mathbf{x} = [x_1, \ldots, x_M]^T$, in the absence of noise, is a convolutive mixture of sources s_1,...,s_L:

$$\mathbf{x}(n) = \sum_{i=1}^{L} s_i(n) \star \mathbf{h}(n, \theta_i) \quad (1)$$

where M is the number of microphones, L is the number of sources, $\mathbf{h}(n, \theta_i)$ is the $M \times 1$ time-domain steering vector corresponding to DOA θ_i of source s_i. The short-time Fourier transform (STFT) representation of the signal model is:

$$\mathbf{X}(m, \omega_k) = \sum_{i}^{L} S_i(m, \omega_k)\mathbf{h}(\omega_k, \theta_i) \quad (2)$$

where m is the time-block index, ω_k is the frequency-bin index ($\omega_k = 2\pi k/N, k = 0, \ldots, N-1$, where N is the FFT size), and $\mathbf{h}(\omega_k, \theta_i)$ is the $M \times 1$ frequency-domain steering vector [4].

3. COHERENCE TEST

The details of the coherence test are described in the work of Mohan *et al.* [4]. The coherence test is applied to the estimated covariance matrix to identify rank-1 TF bins. The true $M \times M$ covariance matrix $\mathbf{R}(m, \omega_k)$ is a linear combination of rank-1 outer products of steering vectors weighted by powers $\sigma_i^2(m, \omega_k)$ of the *i*th source over the (m, ω_k)th TF bin:

$$\mathbf{R}(m, \omega_k) = E[\mathbf{X}(m, \omega_k)\mathbf{X}^{\mathbf{H}}(m, \omega_k)] \quad (3)$$

$$= \sum_{i=1}^{L} \sigma_i^2(m, \omega_k)\mathbf{h}(\omega_k, \theta_i)\mathbf{h}(\omega_k, \theta_i)^{\mathbf{H}} \quad (4)$$

The estimated covariance matrix at TF bin (m, ω_k) is computed from C time-blocks:

$$\hat{\mathbf{R}}(m, \omega_k) = \frac{1}{C} \sum_{l=m-C+1}^{m} \mathbf{X}(l, \omega_k)\mathbf{X}^{\mathbf{H}}(l, \omega_k) \quad (5)$$

Equation (4) shows that if the number of sources constituting the TF bin is larger than or equal to the number of microphones, the covariance matrix is full rank; otherwise the covariance matrix is low rank or poorly conditioned. The coherence test identifies approximately rank-1 TF bins in which only one source s_i ($i \in 1, \ldots, L$) is dominant; thus, the covariance matrix of this bin is approximated as:

$$\tilde{\mathbf{R}}(m, \omega_k) = \sigma_i^2(m, \omega_k)\mathbf{h}(\omega_k, \theta_i)\mathbf{h}(\omega_k, \theta_i)^{\mathbf{H}} \quad (6)$$

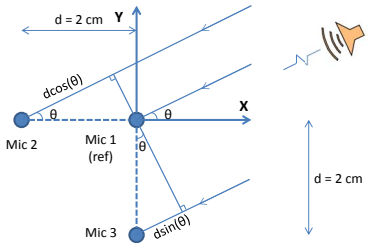

Fig. 2. Microphone array configuration.

4. DOA ESTIMATION ALGORITHM

The block diagram of the DOA estimation algorithm is shown in Fig. 1. Psychoacoustics shows that humans can perceive better sound above the background noise level, and detect the direction in reverberant environments using the precedence effect. The noise-floor tracking and onset detection are inspired by these features of human hearing. These two stages identify TF bins where the signal is substantially larger than noise and the direct signal is dominant over the reflections. Thus the signal model can be approximated as the signal model in a zero-noise and non-reverberant environment as shown in (1). Only the output of the reference microphone is used for noise-floor tracking and onset detection.

Noise-floor tracking: Noise-floor tracking is a popular technique in speech enhancement [7, 8]. In this DOA estimation algorithm, a simple noise-floor tracking in each frequency band is used. TF bins with power above a certain threshold of noise are selected. The noise-floor is adaptively updated during both noise and signal periods. The noise-floor is increased slowly during signal periods and decreased slowly during noise periods.

$$noise\_floor(m, \omega_k) = \alpha \times noise\_floor(m-1, \omega_k) \quad (7)$$

where α is the updating parameter, $\alpha < 1$ during noise periods and $\alpha > 1$ during signal periods.

Onset detection: The onset of a new sound in the TF domain is marked by a sudden rise in energy in some frequency bands. In this work, a novel onset algorithm that tracks the energy peaks in each frequency band is used to detect such rises. The onset threshold is set to the peak value every time an onset is detected and attenuates gradually after that. A TF bin is identified as an onset if it is larger than the onset threshold by a certain amount. Different from the popular spectral flux onset detection [9] that detects one onset across all frequency bands, this algorithm detects onsets in each frequency band independently.

$$\eta(m, \omega_k) = \begin{cases} X(m, \omega_k) & \text{if } X(m, \omega_k) \text{ is onset} \\ \beta \times \eta(m\text{ - }1, \omega_k) & \text{if otherwise.} \end{cases} \quad (8)$$

where β is the decaying parameter, $\beta < 1$.

Table 1. Simulation results: RMS errors (degrees) and failure rates (%) of DOA estimations for various numbers of speech sources with 20 dB SNR and zero-reverberation.

Number of sources		3	4	5	6	7	8
Coherence Test + MUSIC	RMS	**0.64**	**0.91**	**1.07**	**1.17**	1.54	2.04
	% fail	2	3	8	12	52	58
Coherence Test + Eigenvector	RMS	1.20	1.24	1.27	1.30	1.31	1.37
	% fail	**0**	**0**	**0**	4	**4**	28
The Proposed Method	RMS	1.20	1.23	1.24	1.25	**1.30**	**1.31**
	% fail	**0**	**0**	**0**	**0**	8	**24**

Eigenvector clustering and DOA estimation: After noise-tracking and onset detection, the coherence test is applied to select rank-1 TF bins. At this point, most of the covariance matrices can be approximated by (6). From (6), steering vector $\mathbf{h}(\omega_k, \theta_i)$ can be approximated by the largest eigenvector $\mathbf{u}(m, \omega_k)$ of the covariance matrix: $\mathbf{h}(\omega_k, \theta_i) \approx \gamma \mathbf{u}(m, \omega_k)$, where γ is some constant. By clustering these largest eigenvectors, the DOAs can be estimated from the cluster centroids based on the structure of the steering vectors which is dependent on the arrangement of the microphones. To illustrate the concepts of this work, a triangular omnidirectional microphone array as shown in Fig. 2 is used. The steering vector of this microphone array is:

$$\mathbf{h}(\omega_k, \theta_i) = \begin{pmatrix} 1 \\ \exp \frac{-j\omega_k F d \cos(\theta_i)}{c} \\ \exp \frac{-j\omega_k F d \sin(\theta_i)}{c} \end{pmatrix} \quad (9)$$

where F is the sampling rate and c is the speed of sound. The largest eigenvector $\mathbf{u}(m, \omega_k)$ is normalized so that the first element $\bar{u}_1(m, \omega_k) = 1$. The DOA at the (m, ω_k)th bin can be estimated from the second and third element of the eigenvector as follows:

$$\begin{aligned} \cos(\tilde{\theta}_i) &= \frac{c}{\omega_k F d} \angle \bar{u}_2(m, \omega_k) \\ \sin(\tilde{\theta}_i) &= \frac{c}{\omega_k F d} \angle \bar{u}_3(m, \omega_k) \end{aligned} \quad (10)$$

Let $\mathbf{v}(m, \omega_k) = [\cos(\tilde{\theta}_i), \sin(\tilde{\theta}_i)]^T$. The number of sources L is assumed to be known in advance. In variant 1, k-means is used to find L clusters $C_1, \ldots, C_L$ of all the $\mathbf{v}(m, \omega_k)$. Let $\mathbf{c}_i$ be the centroid of cluster C_i; the DOA of source s_i is estimated as $\hat{\theta}_i = \tan^{-1}(\frac{c_{i2}}{c_{i1}})$. In variant 2, the DOA at each TF bin is estimated as $\tilde{\theta}_i = \tan^{-1}(\angle \bar{u}_3(m, \omega_k) / \angle \bar{u}_2(m, \omega_k))$. The DOAs θ_i of all the sources are estimated from the peaks in the one-dimensional histogram of $\tilde{\theta}_i$.

5. RESULTS AND DISCUSSION

The DOA estimation algorithm was implemented in Matlab. The processing parameters were 44.1 kHz sampling rate, 4 s signal duration, 11.61 ms hamming window with 5.8 ms overlap, $N = 512$ fast Fourier transform and 34.8 ms ($C =$ 5 frames) averaging for covariance matrix estimation. The number of sources varied from 3 to 8. All the outputs were an average of 100 trials. Since the distance between 2 microphones was 2 cm, to prevent spatial aliasing, the maximum frequency used to estimate DOAs was $\omega_{k_{max}} = 2\pi c/(2d) \approx 17\pi$ krad ($f_{max} \approx 8.5$ kHz).

Table 2. Simulation results: RMS errors (degrees) and failure rates (%) of DOA estimations for 6 speech sources in various noise levels and zero-reverberation.

SNR (dB)		20	15	10	5	0
Coherence Test + MUSIC	RMS	1.45	2.49	4.00	6.46	NA
	% fail	20	26	76	88	100
Coherence Test + Eigenvector	RMS	1.78	1.77	7.41	29.92	52.91
	% fail	**0**	**3**	**9**	**34**	**52**
The Proposed Method	RMS	**1.22**	**1.27**	**2.62**	**4.72**	**14.45**
	% fail	**0**	**3**	12	50	93

Table 3. Simulation results: RMS errors (degrees) of DOA estimation for 6 speech sources with 20 dB SNR and various reverberation times (using RT_ 60).

RT_ 60 (s)		0.2	0.4	0.6	0.8	1.0	1.2
Coherence Test + Eigenvector	Variant 1	1.23	2.61	4.24	12.59	13.71	36.36
	Variant 2	0.85	**1.85**	16.49	19.71	21.00	33.94
The Proposed Method	Variant 1	0.96	1.95	**2.22**	**3.22**	**3.64**	**19.24**
	Variant 2	**0.76**	2.90	11.12	13.80	17.58	24.01

5.1. Simulation results

The simulation results compared the performance of three DOA estimation methods in various noise and reverberation conditions. Method 1, proposed by [4], combined the coherence test and multiple signal classification (MUSIC) algorithm. Method 2 combined the coherence test and the eigenvector clustering. The proposed method was a combination of the noise-floor tracking, the onset detection and method 2. Method 2 and the proposed method use both Variant 1 (k-means clustering) and Variant 2 (one-dimensional histogram) to obtain DOAs. Continuous, fast-paced speech signals from BBC broadcasts were convolved with simulated room impulse responses [10] and added with white Gaussian noise. The simulated and experimental room-reverberation settings were the same (Fig. 3). In Method 1, the directional spectra obtained by MUSIC at each TB bin were combined across time and frequency. The MUSIC results were compared with the Variant 2 results of Method 2 and the proposed method. Table 1 shows the root mean square (RMS) errors and the failure rates (determined by the frequency that an algorithm fails to estimate all the DOAs) of all the methods when the number of sources varied. In low-noise and zero-reverberation environments and when the number of sources was fewer than 7, Method 1 performed the best. The reason is that the coherence test selects TF bins which mainly comprise one dominant signal; thus, it indirectly selects high signal-to-noise ratio (SNR) TF bins. Method 2 and the proposed method excelled at large

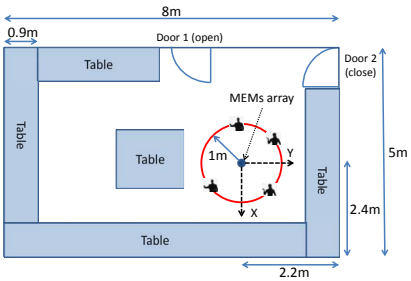

Fig. 3. Experimental setup for MEMs microphone array.

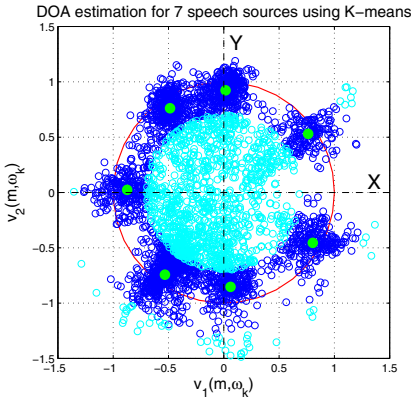

Fig. 4. Variant 1: DOA estimation of the proposed method for 7 speech sources using k-means (1.18 degree RMS error).

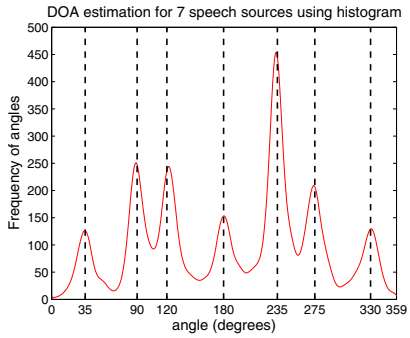

Fig. 5. Variant 2: DOA estimation of the proposed method for 7 speech sources using histogram (1.07 degree RMS error).

Table 4. Experimental results: RMS errors (degrees) and failure rates (%) for various numbers of speech sources.

No of sources			3	4	5	6	7	8
Coherence Test + MUSIC	RMS		2.85	3.64	3.68	4.81	4.89	5.37
	% fail		**0**	**0**	14	30	96	98
The Proposed Method	Variant 1	RMS	1.09	1.63	1.71	2.10	2.85	3.32
		% fail	**0**	**0**	**0**	**0**	**0**	**0**
	Variant 2	RMS	**1.06**	**0.91**	**1.70**	**2.02**	**1.67**	**2.29**
		% fail	**0**	**0**	**0**	**0**	18	66

numbers of sources with lower failure rates and comparable performances. Table 2 shows the RMS errors and failure rates for 6 speech sources in various noise levels and zero-reverberation. When the noise level increased, the eigenvector clustering outperformed the MUSIC algorithm. Table 3 shows the RMS errors for 6 speech sources at 20 dB SNR and various reverberation times. In this simulation, Method 1 returned a 100% failure rate for all cases. With onset detection, the proposed method outperformed Method 2. For Variant 1, taking advantage of the steering-vector structure of microphone array where $\cos^2(\tilde{\theta}_i) + \sin^2(\tilde{\theta}_i) = 1$, only $\mathbf{v}(m, \omega_k)$ that were near to the unit circle were used in k-means clustering to improve the accuracy. Consequently, Variant 1 performed better than Variant 2. In brief, the proposed method surpasses other methods when there are many speech sources and the environment is noisy and heavily reverberant.

5.2. Experimental results

The experiment was conducted in a normal office room during working hours with air-con and other background noise. A MEMS microphone array (with the same arrangement as Fig. 2) was used to record data. The measured sound-pressure level was 62 dB. The average estimated SNR was 19.5 dB. The layout of the room is shown in Fig. 3. The height of the room and the MEMS array were 2.7 m and 1.5 m respectively. In this low-noise and low-reverberant environment, Method 2 and the proposed method performed similarly. Therefore, only the results of Method 1 and the proposed method when the number of sources varied form 3 to 8 were shown in table 4. Method 1 had high failure rates for 6 sources and above. The proposed method estimated DOAs with RMS errors less than 3 degrees. The average unoptimized computation time of Method 1 and the proposed method using an Intel Xeon (R) CPU with 10GB RAM were 119 s and 3.8 s respectively. DOA estimation using the eigenvector clustering outperformed the MUSIC algorithm in both accuracy and computation time. Variant 2 performed better than Variant 1 but with higher failure rates when the number of sources increased. Fig. 4 and Fig. 5 show the well-formed clusters for the k-means and the well-formed peaks for the histogram in estimating DOAs of 7 speech sources. This suggests that the proposed algorithm can be used to estimate the number of speech sources in a mixture by using the elbow method in k-means clustering or counting the number of peaks in the histogram. As a side note, the performance of the proposed method is insensitive to the selection of the threshold values α and β in the noise-floor tracking and the onset detection. In conclusion, the proposed method robustly estimates 8 speech sources (and potentially more) with high accuracy in noisy and reverberant environments using 3 omnidirectional microphones. This method can be applied for other microphone configurations with some changes in the clustering of eigenvectors corresponding to the steering-vector structures.

6. ACKNOWLEDGEMENT

We would like to thank Aaron Jones from Sonistic (http://www.sonistic.com/) for donating the MEMS microphone array that we used to perform our experiments.

7. REFERENCES

[1] Scott Rickard and Ozgiir Yilmaz, "On the approximate w-disjoint orthogonality of speech," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*. IEEE, 2002, vol. 1, pp. I–529.

[2] Shoko Araki, Hiroshi Sawada, Ryo Mukai, and Shoji Makino, "DOA estimation for multiple sparse sources with normalized observation vector clustering," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*. IEEE, 2006, vol. 5.

[3] Yimin Zhang, Weifeng Ma, and Moeness G Amin, "Subspace analysis of spatial time-frequency distribution matrices," *Signal Processing, IEEE Transactions on*, vol. 49, no. 4, pp. 747–759, 2001.

[4] Satish Mohan, Michael E Lockwood, Michael L Kramer, and Douglas L Jones, "Localization of multiple acoustic sources with small arrays using a coherence test," *The Journal of the Acoustical Society of America*, vol. 123, pp. 2136, 2008.

[5] Hans Wallach, Edwin B Newman, and Mark R Rosenzweig, "The precedence effect in sound localization," *The American Journal of Psychology*, vol. 62, no. 3, pp. 315–336, 1949.

[6] Jie Huang, Noboru Ohnishi, and Noboru Sugie, "Sound localization in reverberant environment based on the model of the precedence effect," *Instrumentation and Measurement, IEEE Transactions on*, vol. 46, no. 4, pp. 842–846, 1997.

[7] Yariv Ephraim and David Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 32, no. 6, pp. 1109–1121, 1984.

[8] Rainer Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *Speech and Audio Processing, IEEE Transactions on*, vol. 9, no. 5, pp. 504–512, 2001.

[9] Simon Dixon, "Onset detection revisited," in *Proc. of the Int. Conf. on Digital Audio Effects (DAFx-06)*, 2006, pp. 133–137.

[10] Emanuel AP Habets, "Room impulse response generator," *Technische Universiteit Eindhoven, Tech. Rep*, vol. 2, no. 2.4, pp. 1, 2006.