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 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 direction-of-arrivals of the sources are estimated based on the cluster centroids. Simulation and experimental results show that this method is able to localize more than 8 sources with small errors using only 3 omnidirectional microphones.

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 areas such as surveillance, tracking, tele-conferencing 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 [2] 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. [3] cluster normalized TF bins and estimate the DOAs from the cluster centroids. Zhang et al. [4] 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. [1] 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.

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] showed that humans have such ability because of the precedence effect: when the lag between the first arriving sound and subsequent sounds are sufficiently short (< 40ms), 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 performance for both simulation and experimental data.
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, \( x = [x_1, \ldots, x_M]^T \), in the absence of noise, is a convolutive mixture of sources \( s_1, \ldots, s_L \)

\[
x(n) = \sum_{i=1}^{L} s_i(n) \ast h(n, \theta_i)
\]

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

\[
X(m, \omega_k) = \sum_{i=1}^{L} S_i(m, \omega_k) h(\omega_k, \theta_i)
\]

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

3. COHERENCE TEST

The details of the coherence test are described in the work of Mohan et al. [1]. 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 \((m, \omega_k)\)th TF bin:

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

\[
= \sum_{i=1}^{L} \sigma_i^2(m, \omega_k) h(\omega_k, \theta_i)h(\omega_k, \theta_i)^H
\]

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

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

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

\[
\hat{\mathbf{R}}(m, \omega_k) = \sigma_i^2(m, \omega_k) h(\omega_k, \theta_i)h(\omega_k, \theta_i)^H
\]

4. DOA ALGORITHM

The block diagram of the DOA 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 environment and non-reverberant as shown in (1). Only the output of the reference microphone is used for noise-floor tracking and onset detection.

\[
\text{noise}\_\text{floor}(m, \omega_k) = \alpha \times \text{noise}\_\text{floor}(m - 1, \omega_k)
\]

where \( \alpha \) is the updating parameter, \( \alpha > 1 \) during noise period and \( \alpha < 1 \) during signal period.

\textbf{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.

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

where \( \beta \) 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 20dB SNR and zero-reverberation

<table>
<thead>
<tr>
<th>No of sources</th>
<th>Coherence Test</th>
<th>+ MUSIC</th>
<th>Eigenvector</th>
<th>The Proposed Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMS</td>
<td>% fail</td>
<td>RMS</td>
<td>% fail</td>
</tr>
<tr>
<td>3</td>
<td>0.64</td>
<td>3</td>
<td>1.20</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0.91</td>
<td>5</td>
<td>1.24</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1.07</td>
<td>8</td>
<td>1.27</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>1.17</td>
<td>12</td>
<td>1.30</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>1.54</td>
<td>52</td>
<td>1.31</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>2.04</td>
<td>58</td>
<td>1.37</td>
<td>25</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>SNR (dB)</th>
<th>Coherence Test</th>
<th>+ MUSIC</th>
<th>Eigenvector</th>
<th>The Proposed Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>1.45</td>
<td>20</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>2.49</td>
<td>26</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>4.00</td>
<td>76</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>6.46</td>
<td>88</td>
<td>100</td>
<td>52</td>
</tr>
<tr>
<td>0</td>
<td>NA</td>
<td>100</td>
<td>52</td>
<td>24</td>
</tr>
</tbody>
</table>

Table 3. Simulation results: RMS errors (degrees) of DOA estimation for 6 speech sources with 20dB SNR and various reverberation times (using RT₆₀)

<table>
<thead>
<tr>
<th>RT₆₀ (s)</th>
<th>Coherence Test</th>
<th>+ MUSIC</th>
<th>Eigenvector</th>
<th>The Proposed Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>1.23</td>
<td>2.61</td>
<td>4.24</td>
<td>12.59</td>
</tr>
<tr>
<td>0.4</td>
<td>1.35</td>
<td>1.85</td>
<td>16.49</td>
<td>19.71</td>
</tr>
<tr>
<td>0.6</td>
<td>1.45</td>
<td>1.95</td>
<td>22.32</td>
<td>24.64</td>
</tr>
<tr>
<td>0.8</td>
<td>1.53</td>
<td>2.22</td>
<td>27.32</td>
<td>29.64</td>
</tr>
<tr>
<td>1.0</td>
<td>1.62</td>
<td>2.32</td>
<td>32.32</td>
<td>32.64</td>
</tr>
<tr>
<td>1.2</td>
<td>1.71</td>
<td>2.42</td>
<td>37.32</td>
<td>33.64</td>
</tr>
</tbody>
</table>

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 \( \gamma \) 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

\[
\mathbf{h}(\omega_k, \theta_i) = \begin{pmatrix}
\exp\left(-j\omega_F d \cos(\theta_i)\right) \\
\exp\left(-j\omega_F d \sin(\theta_i)\right)
\end{pmatrix}
\]

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 \( u_1(m, \omega_k) = 1 \). The DOA at the \((m, \omega_k)\) th bin can be estimated from the second and third element of the eigenvector as follows

\[
\begin{align*}
\cos(\hat{\theta}_i) &= \frac{c}{\omega_F F_d} \frac{\mathbf{u}_2(m, \omega_k)}{\mathbf{u}_3(m, \omega_k)} \\
\sin(\hat{\theta}_i) &= \frac{c}{\omega_F F_d} \frac{\mathbf{u}_3(m, \omega_k)}{\mathbf{u}_2(m, \omega_k)}
\end{align*}
\]

Let \( \mathbf{v}(m, \omega_k) = [\cos(\hat{\theta}_i), \sin(\hat{\theta}_i)]^T \). 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 \( c_i \) be the centroid of cluster \( C_i \); the DOA of source \( s_i \) is estimated as \( \hat{\theta}_i = \tan^{-1}\left(\frac{c_i}{\omega_F F_d} \mathbf{u}_2(m, \omega_k) / \mathbf{u}_3(m, \omega_k)\right) \). The DOAs \( \hat{\theta}_i \) for all the sources are estimated from the peaks in the one-dimensional histogram of \( \hat{\theta}_i \).

5. RESULTS AND DISCUSSION

The DOA algorithm was implemented in Matlab. The processing parameters were 44.1kHz sampling rate, 4s signal duration, 11.61ms hamming window with 5.8ms overlap, \( N = 512 \) fast Fourier transform point and 120ms \((C = 20\) frames) averaging for covariance matrix estimate. The number of sources varied from 3 to 8. All the outputs were average of 100 trials. Since the distance between 2 microphones was 2cm, to prevent spatial aliasing, the maximum frequency used to estimate DOAs was \( \omega_{max} = \frac{2\pi c}{2F_d} \approx 17\pi kHz \).

5.1. Simulation results

The simulation results compared the performance of method 1 (combination of the coherence test and multiple signal classification (MUSIC) algorithm [1]) with method 2 (combination of the coherence test and the eigenvector clustering) with the proposed method (combination of the noise-floor tracking, the onset detection, the coherence test and the eigenvector clustering) in various noise and reverberation conditions. Speech signals from BBC broadcasts (continuous, fast-paced speech) were convolved with simulated room impulse responses [10] and added to white Gaussian noise. The simulated room reverberation setting was the same as the experimental setting (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 results of variant 2 using histogram in method 2 and the proposed method. Table 1 shows the root mean square (RMS) errors of all estimation methods and the failure rates (determined by the frequency that an algorithm fails to estimate all the DOAs) when the number of sources varied. In low-noise and non-reverberation environment and when the number of source was fewer than 7, method 1 performed the best. The reason is that the coherence test selects TF bins which is mainly made up by a dominant signal; thus, it indirectly selects high signal to noise ratio (SNR) TF bins. Method 2 and the proposed method excelled at large number of sources with lower failure rates and comparable performance. Table 2
Table 3 shows the RMS errors for 6 speech sources in various noise levels and zero-reverberation. When noise level increased, the eigenvector clustering excelled the MUSIC algorithm. The proposed method outperformed method 2. Variant 1 performed better than variant 2. In brief, the proposed method surpassed other methods in noisy, heavy reverberant environment and with many speech sources.

5.2. Experiment results

The experiment was conducted in a normal office room during working hour 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 in the room was 62dB. 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 2.7m and 1.5m height respectively. In this low-noise and low-reverberant environment experiment, method 2 and the proposed method performed similarly. Table 4 shows the RMS errors and failure rates for method 1 and the proposed method (both variant 1 and 2) when the number of sources varied from 3 to 8. Method 1 had high failure rates for 6 sources and above. The proposed method robustly estimated DOAs with RMS errors less than 3 degrees. The average unoptimized computation time of method 1 and the proposed method using Intel Xeon (R) CPU with 10GB RAM were 119s and 3.8s 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 source increased. Fig. 4 and Fig. 5 show well-formed clusters for both the k-means and the one-dimensional histogram in estimating DOAs of 7 speech sources. For variant 1, taking the advantage of the steering vector structure of microphone array where $\cos(\tilde{\theta}_i) + \sin(\tilde{\theta}_i) = 1$, only $v(m, \omega_k)$ that laid near the unit circle were used in k-means clustering to improve the accuracy. The well-formed clusters in Fig. 4 and the well-formed peaks in Fig. 5 suggest that the proposed algorithm can be used to estimate the number of speech sources in a mixture. As a side note, the performance of the proposed method is insensitive to the selection of the threshold values $\alpha$ and $\beta$ 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 environment using only 3 omnidirectional microphones. This method can be applied for other microphone configurations with some minor changes in the clustering of the eigenvectors corresponding to the steering vector structures.

![Fig. 3. Experimental setup for MEMs microphone array](image1)

![Fig. 4. Variant 1: DOA estimation of the proposed method for 7 speech sources using K-means (1.18 degree RMS error)](image2)

![Fig. 5. Variant 2: DOA estimation of the proposed method for 7 speech sources using histogram (1.07 degree RMS error)](image3)

<table>
<thead>
<tr>
<th>No of sources</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coherence Test + MUSIC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMS</td>
<td>2.85</td>
<td>3.04</td>
<td>3.08</td>
<td>4.81</td>
<td>4.89</td>
<td>5.33</td>
</tr>
<tr>
<td>% fail</td>
<td>0</td>
<td>0</td>
<td>14</td>
<td>30</td>
<td>96</td>
<td>98</td>
</tr>
<tr>
<td>The Proposed Method</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variant 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMS</td>
<td>1.09</td>
<td>1.03</td>
<td>1.14</td>
<td>2.10</td>
<td>2.85</td>
<td>3.32</td>
</tr>
<tr>
<td>% fail</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Variant 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMS</td>
<td>1.06</td>
<td>0.94</td>
<td>1.70</td>
<td>2.02</td>
<td>1.67</td>
<td>2.29</td>
</tr>
<tr>
<td>% fail</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>66</td>
</tr>
</tbody>
</table>
6. ACKNOWLEDGEMENT

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

7. REFERENCES


