Direction-of-Arrival Estimation of Multiple Speakers Using a Planar Array

Dongwen Ying, Ruohua Zhou, Junfeng Li, Jielin Pan, and Yonghong Yan

Key Laboratory of Speech Acoustics and Content Understanding, Chinese Academy of Sciences

yingdongwen@hccl.ioa.ac.cn

Abstract

This paper presents a novel method to estimate the direction-of-arrivals (DOAs) of multiple speakers based on time-frequency sparsity. The acoustic interferences are first suppressed, and then a concave cost function (CCF) is utilized to estimate bin-wise DOAs. The DOAs of sources are subsequently identified by picking peaks in a histogram of bin-wise azimuths. Three aspects distinguish this method from conventional ones. First, a closed-form solution to bin-wise DOA is given by CCF, which replaces the extensively used grid search and enables high computational efficiency. Second, signal enhancement is employed to suppress acoustic interferences on phase spectra. Last, the time-delay weights mitigate the effect of delay outliers. The pair-contributions of this paper are threefold. First, special attention is given to robustness in two aspects. The acoustic interferences are effectively suppressed on phase spectrum. The pairwise time delays are weighted to mitigate the effect of delay outliers on DOA. Few methods consider signal enhancement or delay weights. Second, a TF concave cost function (TF-CCF) is proposed for DOA estimation. This function has a unique extremum and is capable of deriving a closed-form solution to DOA. This capability enables high computational efficiency.

Index Terms: Source localization, eigenanalysis, direction of arrival, time-frequency sparsity, signal enhancement.

1. Introduction

Multiple sound source localization is an indispensable component in numerous systems such as speech separation, speech enhancement, and automatic camera monitoring. Direction-of-arrival (DOA) estimation using microphone arrays is a primary technique that supports these systems. Numerous methods have been proposed in an effort to address DOA estimation. Eigenanalysis [1, 2, 3] and spatial spectral estimation [4, 5, 6] are the most popular techniques. It is well known that their computational efficiency is undesirable due to grid search. A closed-form solution to DOA can drastically improve the computational efficiency. However, a concave/convex cost function that is capable of deriving a closed-form solution is difficult to be designed on the full band. Fortunately, time-frequency (TF) sparsity of speech signals can simplify multiple source localization on the full band into single-source localization at individual bins [7]-[15]. Based on the assumption of sparsity, signal models can be substantially simplified to provide a closed-form solution at each bin. However, the acoustic robustness of sparsity-based methods is undesirable because bin-wise DOA estimation is affected by acoustic interferences. In addition, bin-wise estimation is likely to suffer from spatial aliasing, which is encountered when time delays are bin-wise estimated with widely spaced microphones. Limiting inter-microphone spaces will cause performance degradation.

This paper proposes a novel method on the basis of TF sparsity. This method can be regarded as an extension of single source localization [16] to multiple source localization. The contributions of this paper are threefold. First, special attention is given to robustness in two aspects. The acoustic interferences are effectively suppressed on phase spectrum. The pairwise time delays are weighted to mitigate the effect of delay outliers on DOA. Few methods consider signal enhancement or delay weights. Second, a TF concave cost function (TF-CCF) is proposed for DOA estimation. This function has a unique extremum and is capable of deriving a closed-form solution to DOA. This capability enables high computational efficiency.

Last, a spatial de-aliasing method is proposed. In theory, the microphone spacing is no longer restricted by the sampling frequency of input signals.

2. Problem Formulation

Speech sources have a special property in TF domain. When a recording contains multiple speakers, there are segments of time in which only one speaker is active and the remaining speakers are inactive in the state of short pauses or silent breaks. Even though more than one speaker is active in some segments, the signal power of different speakers in the TF domain may occupy different sets of discrete frequencies. At a given TF bin, there is high likelihood that at most one source is dominating in power and the contributions from the remaining sources are negligible. A simplified model is presented based on TF sparsity.

Let us consider $D$ sparse sources that impinge on a $K$-element planar array. Assuming that the size of the array aperture is small relative to the distance from the sources to the array, a plane wave propagation model is adopted in a far-field scenario. Using the vector notation, the model is written as

$$\mathbf{x}(\omega_i) = \mathbf{a}_d(\omega_i) S_d(\omega_i) + \mathbf{n}(\omega_i) \quad (1)$$

where

$$\mathbf{x}(\omega_i) = [X_1(\omega_i), \cdots, X_K(\omega_i)]^T,$$

$$\mathbf{a}_d(\omega_i) = [A_1(\omega_i), \cdots, A_K(\omega_i)]^T,$$

$$\mathbf{n}(\omega_i) = [N_1(\omega_i), \cdots, N_K(\omega_i)]^T,$$

$$A_{k,d}(\omega_i) = e^{-j\omega_i \psi_{k,d}},$$

where $d$ denotes the dominant source, $(\cdot)^T$ denotes the transpose, $0 \leq \omega_i \leq 2\pi$ is the digital frequency at the $i$th bin, and $j = \sqrt{-1}$ is the imaginary unit. The symbol $S_d(\omega_i)$ denotes the spectrum of the dominant source. The noise term $N_k(\omega_i)$ represents the sum of all spectra, except spectra of the direct sounds from $D$ sources. Both the environmental noise and the reverberation are included in $N_k(\omega_i)$. The DOAs of sources are expressed by the propagation times $\psi_{k,d}$. This model simplifies broadband multiple source localization into narrowband single source localization.

3. Proposed method

This section presents a three-stage method for DOA estimation, as shown in Fig. 1. The details are provided as follows.
3.1. Signal enhancement based on eigenanalysis

The acoustic interferences such as reverberation and noise are the major hindrance to DOA estimation. An eigenanalysis-based approach is utilized to suppress the acoustic interferences. This enhancement method does not require the array geometry knowledge or DOA. It serves as a self-organized beamformer.

This enhancement utilizes a correlation matrix defined as

$$R(\omega_i) = E[x(\omega_i)x^H(\omega_i)],$$  

(2)

where $(\cdot)^H$ and $E(\cdot)$ denote the conjugate transpose and expectation over time, respectively. The enhancement rests on an assumption that the directional component is uncorrelated with the less-directional component. By substituting Eq. (1) into Eq. (2), the correlation matrix is written as

$$R(\omega_i) = n_2(\omega_i)n_2^H(\omega_i)E[|s_d(\omega_i)|^2] + N(\omega_i),$$  

(3)

where

$$N(\omega_i) = E[n(\omega_i)n^H(\omega_i)].$$  

(4)

Although the early reverberation in less-directional components may be correlated with directional components, such correlation is often small relative to the correlation between directional components. So, this assumption is valid in a practical sense.

By eigenvalue decomposition of $R(\omega_i)$, we obtain

$$R(\omega_i) = U(\omega_i)A_iU^H(\omega_i),$$  

(5)

where

$$U(\omega_i) = [u_1(\omega_i), \ldots, u_K(\omega_i)],$$

$$A_i = diag(\lambda_1, \ldots, \lambda_i, K),$$

where $U(\omega_i)$ is the eigenvector matrix, and $A_i$ is the eigenvalue matrix, in which the eigenvalues are sorted in a descending order. Assuming that less-directional components are spatially white, i.e., $N(\omega_i) = \kappa_iI$, and its intensity is much smaller than the intensity of directional components, i.e., $|a_i(\omega_i)s_d(\omega_i)|^2/K >> \kappa_i$, the principal eigenvector can be approximated as the steering vector as follows:

$$u_1(\omega_i) \approx e^{-j\omega_i z_1}\|a_i(\omega_i)|,$$

$$u_{1+i}(\omega_i) \approx e^{-j\omega_i(z_k+i)},$$  

(6)

where $z_1$ is an arbitrary real constant that is introduced by the complex eigenvalue decomposition. This formula indicates that the principal components of observed signals can be regarded as directional eigenvectors. This enhancement can mitigate the effects of acoustic interferences on the phase spectra and so improve the robustness of sound source localization. It can be taken as an extension of the subspace enhancement algorithm [17] into sparse signals.

Without considering spatial aliasing, the pairwise time delay from the $i$th to $j$th microphone is derived from inter-microphone phase differences of the principal components as

$$\tau_{m,i} = \frac{\angle u_{1}(\omega_i) - \angle u_{1}(\omega_j)}{\omega_i}$$  

(7)

where $\angle(\cdot)$ denotes the phase operation, $\tau_{m,i}$ denotes time difference of arrival for the dominant source at the $i$th bin, and $m$ denotes the index of a microphone pair. For a given bin, a total of $M = K(K - 1)/2$ pairwise time delays exist across all microphones.

3.2. Weighted concave cost function

Bin-wise DOAs are estimated from the time delays and the array topology. The estimation procedure is based on a cost function. To achieve high computational efficiency, this function is required to be concave or convex. In addition, the function should consider robustness because noise and reverberation exist in the principal components. A cost function is presented based on the following considerations.

The array topology is expressed by a group of unit vectors $\{g_m = [g_{m,1}, g_{m,2}, 0]^T | m = 1, \cdots, M\}$; each unit vector denotes the direction from one microphone to another microphone. The third dimension, which is set to zero, signifies that all microphones are arranged in a plane. The delay $\tau_{m,i}$ determines the included angle $\theta_{m,i}$ between the $m$th microphone pair direction and the impinging direction of the dominant source, the cosine of which is defined as

$$\cos \theta_{m,i} = c\tau_{m,i}/r_m,$$  

(8)

where $r_m$ denotes the microphone distance of the $m$th pair, and $c$ is the sound speed. By utilizing the geometric relationship, this cosine function can also be represented as

$$\cos \theta_{m,i} = e_{m,i}^T g_m \gamma_i,$$  

(9)

where $\gamma_i = [\gamma_{1,i}, \gamma_{2,i}, \gamma_{3,i}]^T$ is a unit direction vector that represents the DOA of dominant sources at the $i$th bin.

Under desirable acoustic conditions, $\cos \theta_{m,i}$ is infinitely close to $\cos \hat{\theta}_{m,i}$, which is described as

$$\varepsilon_i(\gamma) = \sum_{m=1}^{M} w_{m,i}(g_m^T g_m - c\tau_{m,i}/r_m)^2,$$  

(10)

where $w_{m,i}$ is a coefficient that weights the reliability of the $m$th time delay. The DOA is estimated by minimizing the error as follows:

$$\hat{\gamma}_i = \min_{\gamma} \varepsilon_i(\gamma)$$  

subjected to: $\gamma^T \gamma = 1.$

The optimal estimator based on Eq. (11) is constructed using the Kuhn-Tucker necessary conditions for constrained minimization. The gradient Lagrangian equation is described as

$$L(\gamma, \mu) = \varepsilon_i(\gamma) + \mu(\gamma^T \gamma - 1),$$  

(12)

where $\mu$ is the Lagrangian multiplier. Eq. (12) can be confirmed to be concave. From the derivative function $\nabla_{\gamma} L(\gamma, \mu) = 0$, we obtain

$$\left( \hat{\gamma}_{1,i}, \hat{\gamma}_{2,i} \right) = \left( \sum_{m=1}^{M} w_{m,i} g_m^T g_m \right)^{-1} \sum_{m=1}^{M} c w_{m,i} \hat{\tau}_{m,i} g_m^T / r_m,$$  

(13)
where \( g_m' = [g_{m,1}, g_{m,2}]^T \), and \( w_{m,i} \) is considered a constant in the solution of DOA. Eq. (13) explicitly represents the DOA as a linear function of all pairwise delays. The azimuth and elevation are obtained from the DOA \( \hat{\gamma}_i \) as follows:

\[
\alpha_i = \begin{cases} 
\arccos(\gamma_{1,i}/\sqrt{\gamma_{1,i}^2 + \gamma_{2,i}^2}) & \text{if } \gamma_{2,i} \geq 0 \\
\pi + \arccos(\gamma_{1,i}/\sqrt{\gamma_{1,i}^2 + \gamma_{2,i}^2}) & \text{if } \gamma_{2,i} < 0,
\end{cases}
\]

Each delay is weighted by the error between \( \theta_{m,i} \) and \( \hat{\theta}_{m,i} \):

\[
\delta_{m,i} = \arccos(g_m^T \hat{\gamma}_i) - \arccos(c\tau_{m,i}/r_m),
\]

where \( g_m \) is the inverse cosine operation. Assuming that the error confirms a zero-mean Gaussian distribution with variance \( \sigma_i^2 = \sum_{m=1}^{M} \delta_{m,i}/M \), the normalized weight of each delay is derived from the likelihood as

\[
w_{m,i} = \exp(-\delta_{m,i}^2/\sigma_i^2)/\sum_{m=1}^{M} \exp(-\delta_{m,i}^2/\sigma_i^2).
\]

The delay outliers are usually associated with larger errors and small weights, which reduces the effect of the outliers.

### 3.3. DOA estimation under spatial aliasing

Because bin-wise DOAs and weight coefficients are interdependent, an iterative algorithm is presented for their solutions. An initial DOA is selected from a data set and is subsequently utilized to calculate new weights. A new DOA is estimated from the new weights. This iterative procedure continues until the DOA converges. Phase unwrapping is incorporated into this iteration. In spatial aliasing, the set of all potential delays for a given DOA is defined as

\[
B_{m,i} = \{ \tau_{m,i} \mid \tau_{m,i} = \frac{\angle u_{13}(\omega_i) - \angle u_{13}(\omega_i) + 2\pi l}{\omega_i}, \eta_m < \tau_{m,i} < \eta_m \},
\]

where \( \eta_m \) denotes the maximal delay determined by microphone spacing, i.e. \( \eta_m = r_m/c \). The idea of phase unwrapping is the selection of a delay in the set with the smallest distance to a given DOA, which is described as

\[
\tau_{m,i} = \arg \min_{r \in B_{m,i}} |g_m^T \hat{\gamma}_i r_m - c\tau|,
\]

where \( \hat{\gamma}_i \) is a given DOA in the iterative procedure.

The selection of an initial DOA closest to the real target enables the iteration to converge toward the real target. We construct a set of initial DOAs, \( \{\chi_1, \cdots, \chi_M\} \), which is uniformly distributed in the \( 360 \times 90^\circ \) grid space. The distance from a given candidate to a pairwise delay is given by

\[
b_i(\chi_h) = \sum_{m=1}^{M} \min_{r \in B_{m,i}} \left[ (\chi_h^T g_m r_m - c\tau)^2 \right],
\]

where \% denotes a floating-point remainder operation, and \( T_i \) denotes the period at the \( i \)th bin, given by

\[
T_i = 2\pi/\omega_i.
\]

The candidate direction with the smallest distance is taken as the initial DOA. The estimation procedure is summarized in Algorithm 1, in which \( \epsilon \) is a constant greater than but close to zero.

### Algorithm 1 DOA Estimation at a Frequency Bin

**Require:** FFT coefficients at the \( i \)th bin, \( \{x_1(\omega_i), \cdots, x_K(\omega_i)\} \).

**Ensure:** Azimuth of the dominant source at the \( i \)th bin, \( \hat{\alpha}_i \).
1. Perform eigenanalysis-based enhancement;
2. Calculate the delay sets \( \{B_1, \cdots, B_M\} \) using Eq. (17);
3. Chose an initial DOA \( \hat{\gamma}_i \) from \( \{\chi_1, \cdots, \chi_M\} \) using Eq. (19);
4. repeat
5. Unwrap all inter-phase differences using Eq. (18);
6. Let \( p = \hat{\gamma}_i \), and calculate the new weights using Eqs. (15) and (16);
7. Re-calculate \( \hat{\gamma}_i \) with the new weights using Eq. (13);
8. until \( (1 - p^T \hat{\gamma}_i < \epsilon) \)
9. Calculate azimuth \( \hat{\alpha}_i \) from \( \hat{\gamma}_i \) via Eq. (14).

Because the azimuth discrimination is much more precise than the elevation discrimination for planar arrays, the azimuth is utilized to identify the sources. The histogram of the bin-wise azimuths is weighted by normalized eigenvalues \( \lambda_i \). The occurrence of a histogram bin is calculated as the summation of all principal eigenvalues with azimuths assigned to this bin. Spurious peaks in the histogram are smoothed out by a Hanning window. Let \( \bar{b} \) denote the window function of length \( 2\epsilon + 1 \); that is, \( \sum_{r=-\epsilon}^{\epsilon} b(r) = 1 \). The smoothed occurrence is provided by

\[
\phi(\alpha) = \sum_{r=-\epsilon}^{\epsilon} \bar{b}(r) \sum_{\alpha_i \in [r-\delta/2, r+\delta/2]} \lambda_i,
\]

where \( \delta \) denotes the width of a histogram bin and \( \alpha \) denotes the discretized azimuth.

The sound sources are identified by picking peaks in the histogram. Each source corresponds to a peak with occurrence greater than the threshold \( \Delta \). The threshold is set as

\[
\Delta = O_{avg} + g(O_{max} - O_{avg}),
\]

where \( O_{avg} \) and \( O_{max} \) denote the average and maximum of the smoothed occurrence, respectively, and the coefficient \( g \) (\( 0 < g < 1 \)) is set by experience. The number of sources are determined by counting the number of distinct peaks.

### 4. Evaluation

The proposed method was tested in a simulated room with dimensions of \( 6 \times 8 \times 3 \) m in far-field scenarios. The microphone array was placed at the center of the room. The circular radius was 0.08 m in the following experiments. The speech sources were located at a horizontal distance of 0.9 m from the array center. The vertical heights of the speech sources and the array were 0.8 m and 1.5 m, respectively. Since precise discrimination of the elevations is incapable of being provided by the planar array, the evaluation focused on the arrival azimuths. A moderately adverse environment was simulated using the image source method [18]. The reverberation time is set as 300 ms. The white noise was artificially added to the simulated signals with SNR=10 dB.

Monosyllabic utterances from different speakers were artificially mixed with equal intensity. The energy maxima of different speakers’ utterances temporally coincided to form a mixture segment. For simplicity, it is assumed that all speakers’ utterances were simultaneously present at every interval that contained speech signal. The total number of sources is three times
the number of tested intervals. The signal length is 39 seconds. The signal sampling rate is 8000 Hz.

The proposed method was compared with STMV [5] and TF-CHB [13]. TF-CCF is our proposed method. TF-CHB and TF-CCF are sparsity-based methods, in which the azimuths are estimated at individual bins and summarized across all bins. TF-CHB and STMV determine the number of sources in a way similar to the proposed method. They each employed 32-ms frames with a half frame overlap, and estimated DOAs using a 7-frame sliding window. STMV and TF-CHB performed hypothesis tests at 1-degree intervals.

The first simulated experiment investigated the spatial resolution of three methods. Two speakers were located at various azimuth spaces. The coefficient that tunes the threshold was set to \( g = 0.2 \) for all methods. Fig. 2 illustrates the histograms for different spacings of azimuth angles. This figure illustrates that TF-CCF achieved the best performance among the three methods even when the angle space was enlarged to 44°. The spatial resolution of TF-CHB is similar to the spatial resolution of TF-CCF. STMV is incapable of distinguishing two speakers when their azimuth spacing is less than 14°. The high resolution of sparsity-based methods is attributed to bin-wise DOA estimation. Because all frequency bins are equally considered in the histogram, the weak sources are not likely to be masked by the strong sources. In contrast, STMV is conducted based on a covariance matrix that is averaged over all bins. The weak frequency components are frequently masked by strong frequency components in the averaging procedure. Especially for highly neighboring sources, the averaging procedure may confuse DOAs. But this averaging operation can smooth out the weak imaginary sources. Sparsity-based methods, such as TF-CHB, will exhibit high false alarms because acoustic interferences are not well suppressed. The reason is that acoustic interferences significantly deteriorate bin-wise DOA estimation [19] and thus yield imaginary sources that induce false alarms. TF-CCF can adequately address acoustic interferences via signal enhancement and delay weighting.

The second experiment investigated the influences of signal enhancement and weighted iteration on performance. The speakers located at the azimuths of 74°, 309.6°, and 352.7°. Without signal enhancement, the pairwise delays were directly obtained from the inter-phase differences of Fourier coefficients. A median filter was utilized to smooth delays over 7 frames. Fig. 3 compares a situation with enhancement and a situation without enhancement. This figure illustrates that the acoustic interferences cause the estimate to deviate from the true direction. The deviation is significantly reduced through enhancement. Fig. 4 illustrates the convergence error that is averaged over all times and frequencies. This figure confirms the contribution of delay weights to acoustic robustness.

The computational load of TF-CCF and STMV were compared using a desktop computer. Both methods were implemented using standard C program. The cosine and sine operations in both methods were realized by table look-up operations. The complex eigenvalue decomposition in TF-CCF was implemented by the CLAPACK package [20]. The experiment shows that TF-CCF runs about 200 times faster than STMV. Since TF-CHB employs grid search instead of the closed-form solutions, TF-CCF outperforms TF-CHB in computational efficiency.

5. Conclusions

In this paper, a novel method for detecting multiple speech sources using a planar array has been proposed. Each stage of this method is conducted based on TF sparsity. Our proposed method possesses the advantages in both computational efficiency and acoustic robustness. Additionally, the proposed method can be applied to not only circular arrays, but also planar arrays with any topology. Because of its capability of spatial de-aliasing, the inter-microphone spacing can exceed the threshold that produces spatial aliasing. These advantages are valuable in real applications.

6. Acknowledgements

The authors would like to thank Dr. M. Cobos for sharing their source code. This work was supported by National Program on Key Basic Research Project (2013CB329302), NSF of China (No. 61271426, 11161140319, and 91120001), Strategic Priority Research Program of CAS (XDA06030100 and XDA06030500), National 863 Program (2012AA012503) and by CAS Priority Deployment Project (KGZD-EW-103-2).
7. References


