ROBUST MULTIPLE SPEECH SOURCE LOCALIZATION USING TIME DELAY HISTOGRAM

Zhaoqiong Huang, Ge Zhan, Dongwen Ying, and Yonghong Yan

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

ABSTRACT

Spatial aliasing and spatial resolution are the two issues faced by most multiple speech source localization methods. The histogram of time delays is a simple but effective method to deal with these two issues on linear arrays. But few methods were capable of applying the time delay histogram to directional-of-arrivals (DOAs) estimation using a planar array. This paper proposes a novel method to estimate DOAs of multiple speech sources based on time delay histograms across all microphones of a planar array. The pairwise time delays of different sources are firstly obtained from each time delay histogram, and then, the time delays are identified with variant speech sources. Eventually, the DOA of each source is estimated by regression over its associated time delays. We conducted some experiments in both simulated and real environments to evaluate the proposed method using an eight-element circular array. The experimental results confirmed not only its high computational efficiency, but also its superiority in spatial resolution and spatial anti-aliasing.

Index Terms— Speech source localization, time delay histogram, spatial aliasing, spatial resolution, direction of arrival.

1. INTRODUCTION

Multiple speech source localization is widely used in numerous applications such as speech enhancement, speech separation, and distant speech recognition [1]. The assumption of speech sparse distribution in the time-frequency (TF) domain is frequently utilized to localize multiple speech sources, and many sparsity-based methods were presented in the past several decades [2] - [6]. Besides the acoustic robustness, the spatial aliasing and spatial resolution are the two challenging issues that are faced by the sparsity-based methods. Speech is a wide-band signal, and spatial aliasing occurs at some high frequencies when the microphones are widely spaced, where one given time delay corresponds to multiple time delay candidates. Limiting the inter-microphone space was often used to avoid spatial aliasing [6], [7], but the small space will degrade the spatial resolution [8]. Traversing all the potential numbers of aliasing period is another method to resolve spatial aliasing for linear arrays [3], [9]. However, the situation is extremely complex for planar arrays, where there may exist a large number of combinations of aliasing period across all microphone pairs in high frequencies. So the traversing method will lead to heavy computational load for planar arrays. A closed-form method of spatial de-aliasing for multiple speech source localization has been presented for real-time speech source localization [10]. But this method can not well treat serious spatial aliasing. The other challenging issue is spatial resolution. When the spatial locations of speech sources are close to each other, the speech sources are difficult to be discriminated and the miss or false detections are likely to occur.

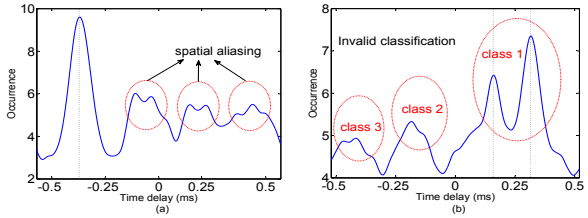

Fig. 1: Typical time delay histograms: (a) Histogram of one speech source with serious spatial aliasing at high frequencies; (b) Histogram of two closely located sources. The dotted lines denote the true time delays.

The histogram analysis is a simple but effective approach to deal with these two issues. Because the periods at variant frequencies are different, the peaks of aliased time delays are not so significant as the peak of actual time delays, as shown in Fig. 1(a). In the other word, the time delay histogram has the capability of spatial anti-aliasing. Moreover, the histogram approach outperforms conventional methods in discrimination of closely located sources. The cluster-based methods were conventionally utilized to identify speech sources [2], [3], [7]. But those methods are likely to confuse two closely located sources as one cluster, and the imaginary sources with low occurrence were usually taken as a real source. On contrast, the histogram method is capable of discriminating the imaginary sources from real sources according to occurrence and telling two closely located sources apart, as illustrated in Fig. 1(b).

The proposed method takes advantage of the time delay histogram to estimate DOAs of multiple speech sources using a planar array. The time delays of each microphone pair are obtained by picking the peaks of the corresponding histogram of time delays at all times and all frequencies. The critical problem is to identify each time delay with a source. An algorithm is presented to identify those time delays with each source. Eventually, the DOA of each source is estimated by means of regression over their associated time delays.

2. PROBLEM FORMULATION

Let us consider D speech sources that impinge on a K-element planar array in a far-field scenario. It is assumed that the size of the array aperture is small relative to the distance from the sources to the array. Speech signal has been shown to be sparsely distributed in the TF domain [11]. At a given TF bin, there is high likelihood

This work was supported by the National Program on Key Basic Research Project (2013CB329302), the National Natural Science Foundation of China (Nos. 61271426, 11461141004, 91120001), the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant Nos. XDA06030100, XDA06030500), and by the CAS Priority Deployment Project (KGZD-EW-103-2).

that at most one speech source is dominating in power and the contributions from the remaining sources are negligible. Based on these assumptions, the signal received by the kth microphone is represented in frequency domain as

$$Y_k(\omega_f) = e^{-j\omega_f \varphi_{k,d}} S_d(\omega_f) + N_k(\omega_f), f \in \{1, \ldots, F\}, \quad (1)$$

where

$$d = \arg \max_{d \in [1:D]} \left| S_d(\omega_f) \right|, \quad (2)$$

where k denotes the microphone index, $0 \leq \omega_f \leq 2\pi$ denotes the digital frequency, f denotes the frequency index, $j = \sqrt{-1}$ denotes the imaginary unit, $\varphi_{k,d}$ denotes the propagation time from the dth source to the kth microphone, $S_d(\omega_f)$ denotes the signal emitted from the dth source, F denotes half short-term Fourier transform (STFT) length, and $N_k(\omega_f)$ denotes the acoustic interferences that comprise the additive noise and reverberation.

There are in total $M = K(K-1)/2$ microphone pairs. For a given TF bin, the time delay is determined by the dominant source. The mth pairwise time delay between the pth and qth microphone can be expressed as

$$\begin{aligned} \widehat{\tau}_{m,f} &= \left[\angle Y_p(\omega_f) - \angle Y_q(\omega_f) \right] / \omega_f \\ &= \varphi_{q,d} - \varphi_{p,d} + n_{m,f} T_f + \xi(\omega_f) \\ &= \phi_{m,f} + n_{m,f} T_f, \end{aligned} \quad (3)$$

where

$$\phi_{m,f} \in [-T_f/2, T_f/2], \quad T_f = 2\pi/\omega_f,$$

where $\angle(.)$ denotes the phase operation, $\xi(\omega_f)$ is the perturbation caused by acoustic interferences, T_f denotes the period at the fth frequency, $\phi_{m,f}$ denotes the temporal phase that is derived from the observed phase difference, and the integer $n_{m,f}$ denotes the number of aliasing periods. $n_{m,f}$ may have several candidates for widely spaced microphones, which leads to several candidates for each time delay. The potential time delays are given by a set:

$$\begin{aligned} B_{m,f} = \Big\{ \tau \Big| & \tau = \phi_{m,f} + n_{m,f} T_f, \\ & \frac{-r_m/c - \phi_{m,f}}{T_f} \leq n_{m,f} \leq \frac{r_m/c - \phi_{m,f}}{T_f} \Big\}, \end{aligned} \quad (4)$$

where c denotes the sound velocity and r_m denotes the distance between the mth microphone pair. The cardinality $|B_{m,f}|$ may be different from TF bin to TF bin. If $|B_{m,f}| > 1$, spatial aliasing occurs at this TF bin. $|B_{m,f}| = 1$ indicates that there is no spatial aliasing. If $|B_{m,f}| = 0$, the time delay at this TF bin is invalid and it will be disregarded in the following processing. For the mth microphone pair, the set of time delay candidates at all frequencies is given by

$$\Gamma_m = \left\{ B_{m,2}, \cdots, B_{m,F} \right\}, \quad (5)$$

where the first frequency is disregarded since it does not contain the information of time delay. The time delays of microphone pairs are obtained by applying the histogram analysis on set $\left\{ \Gamma_1, \ldots, \Gamma_M \right\}$, and then, the DOAs are derived from these delays.

3. PROPOSED METHOD

The basic idea of the proposed method is to estimate time delays of each microphone pair and identify the individual time delay with a speech source. By constructing the time delay histogram of the

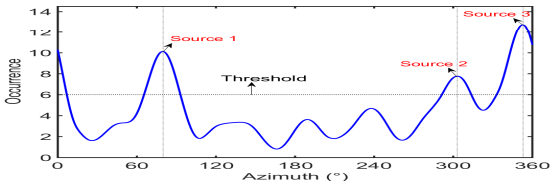

Fig. 2: Azimuth histogram of three speech sources. The dotted lines denote the initial azimuths.

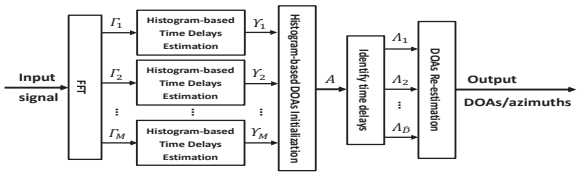

Fig. 3: Block diagram of the proposed method.

mth microphone pair using Γ_m, each significant peak with high occurrence is identified as the time delay of a speech source, which is represented by a set:

$$\Upsilon_m = \left\{ \widehat{\tau}_{m,1}, \cdots, \widehat{\tau}_{m,I_m} \right\}, \quad (6)$$

where I_m denotes the number of distinct peaks. In desirable acoustic conditions, I_m is equal to the source number D. Under adverse environments, however, I_m may be greater than or less than the number of real speech sources. The DOAs are estimated from the set $\Psi = \left\{ \Upsilon_1, \ldots, \Upsilon_M \right\}$.

Generally speaking, the azimuth is the most reliable feature to discriminate different sources for the horizontally placed planar array, and it is therefore utilized to identify time delays in Ψ with speech sources in the proposed method. For a planar array, every two delays (τ_1, τ_2) can determine an azimuth, which is given by

$$\widehat{\alpha}_{\tau_1,\tau_2} = \mathcal{G}(\tau_1, \tau_2), \quad \tau_1 \neq \tau_2, \quad (7)$$

where $\mathcal{G}(\cdot)$ is a regression function that is determined by the array topology, the detail of which is given in reference [12]. There are three cases for a pair of time delays in (7). The first case is that the two delays are associated with the same speech source, where the determined azimuth is often close to the actual azimuth of this source. The second case is that the two delays belong to different sources, where the function $\mathcal{G}$ may have no output or the outputs are randomly distributed. The first two cases only consider the delays correspond to different microphone pairs. The third case is that the two delays correspond to the same microphone pair where the function has no output. If the histogram is constructed on all potential azimuths, each significant peak of occurrence usually corresponds to a speech source, as shown in Fig. 2. These azimuths are taken as the initial estimates, which are given by

$$A = \left\{ \widehat{\alpha}_1, \ldots, \widehat{\alpha}_{\widehat{D}} \right\}. \quad (8)$$

The number of speech sources, $\widehat{D}$, is determined by counting the significant peaks in the azimuth histogram. Each initial azimuth of a source is used to identify time delays in Ψ with this source, and then, this azimuth is refined by regression over all time delays that associated with this source. The time delays are identified by a voting process. Let's define an azimuth set that is associated with a time

delay $\tau_{m,i}$ in Ψ, which is given by

$$\Phi^{(\tau_{m,i})} = \Big\{\alpha | \alpha = \mathcal{G}(\tau, \tau_{m,i}),\ \tau \in \Psi\ \ and\ \ \tau \neq \tau_{m,i}\Big\}. \quad (9)$$

For each $\tau_{m,i}$,

$$\begin{aligned} v_d(\tau_{m,i}) &= v_d(\tau_{m,i}) + 1, \\ if : \mathcal{F}(\alpha - \widehat{\alpha}_d) &< \delta,\ \alpha \in \Phi^{(\tau_{m,i})}, \end{aligned} \quad (10)$$

where $v_d(\tau_{m,i})$ denotes the votes that associate $\tau_{m,i}$ with the dth source, δ denotes a voting threshold, and the minus operation for angular degree is defined as

$$\mathcal{F}(\alpha) = \big|\alpha + 360^\circ \times \widehat{h}\big|, \quad (11)$$

where

$$\widehat{h} = \arg\min_h \big|\alpha + 360^\circ \times h\big|,$$

where h is an integer that minimizes the absolute error. Each time delay is identified to the speech source with the maximal votes, which is given by

$$d(\tau_{m,i}) = \arg \max_{d \in \{1,\ldots,\widehat{D}\}} v_d(\tau_{m,i}). \quad (12)$$

The time delays associated with a speech source are described by

$$\Lambda_d = \Big\{\tau_{d,1}, \cdots, \tau_{d,L_d}\Big\}. \quad (13)$$

It should be noted that some time delays in set Λ_d may correspond to the same microphone pair. In this case, the time delay with the highest votes among these conflicted delays is remained and the other conflicted delays are removed from this set. Afterwards, the time delays identified to the dth speech source is obtained as (13), where L_d is the total number of time delays associated with the dth source. It should be mentioned that $\sum_{d=1}^{\widehat{D}} L_d <= \sum_{m=1}^{M} I_m$ because there may exist some miss detected time delays, i.e., some time delays are not identified to any source. By regression over time delays in Λ_d, the closed-form solution to the azimuth of the dth speech source is given by

$$\widehat{\alpha}_d = \mathcal{G}(\tau_{d,1}, \cdots, \tau_{d,L_d}). \quad (14)$$

4. IMPLEMENTATION

The block diagram of the proposed method is shown in Fig. 3, where the histogram analysis has been used twice. One is to estimate the pairwise time delays, and the other is to estimate the initial azimuths. Spurious peaks in the histograms are smoothed out by a Hanning window. Here, each significant peak is defined as one with occurrence greater than threshold $\triangle$, which is given by

$$\triangle = O_{avg} + \eta(O_{max} - O_{avg}), \quad (15)$$

where O_{avg} and O_{max} denote the average and maximum of the smoothed occurrence, respectively, and the coefficient η $(0 < \eta < 1)$ is set by experience. The estimation is summarized in Algorithm 1, where an algorithm to identify time delays with speech sources is shown as Algorithm 2.

Algorithm 1 : DOAs estimation

1: Calculate time delay candidates at all frequencies using (3), (4) and (5).
2: Construct the time delay histogram for each microphone pair.
3: Estimate the pairwise time delays from histograms and construct the time delay set $\Psi = \Big\{\Upsilon_1, \ldots, \Upsilon_M\Big\}$.
4: Calculate the azimuths of every two time delays in Ψ using (7).
5: Construct the azimuth histogram and determine the number of speech sources $\widehat{D}$ and the initial azimuths $A = \Big\{\widehat{\alpha}_1, \ldots, \widehat{\alpha}_{\widehat{D}}\Big\}$.
6: Identify each time delay in Ψ with a source using Algorithm. 2.
7: Obtain $\Big\{\Lambda_1, \cdots, \Lambda_{\widehat{D}}\Big\}$ and calculate the azimuths of speech sources by regression using (14).

Algorithm 2 : Time delays identification

for each $d \in \{1, \ldots, \widehat{D}\}$ **do**
    $\Lambda_d = \emptyset$.
end for
for each $\tau_{m,i} \in \Psi$ **do**
    for each $d \in \{1, \ldots, \widehat{D}\}$ **do**
        $v_d(\tau_{m,i}) = \emptyset$.
    end for
    for each $\tau \in \Psi$ and $\tau \neq \tau_{m,i}$ **do**
        $\alpha = \mathcal{G}(\tau, \tau_{m,i})$.
        for each $d \in \{1, \ldots, \widehat{D}\}$ **do**
            if $\mathcal{F}(\alpha - \widehat{\alpha}_d) < \delta$ **then**
                $v_d(\tau_{m,i}) = v_d(\tau_{m,i}) + 1$.
            end if
        end for
    end for
    $d(\tau_{m,i}) = \arg \max_{d \in \{1,\ldots,\widehat{D}\}} v_d(\tau_{m,i})$.
    $\Lambda_{d(\tau_{m,i})} = \Lambda_{d(\tau_{m,i})} \bigcup \Big\{\tau_{m,i}\Big\}$.
end for

5. EVALUATION

This section evaluates the proposed method by the simulated and real environments. The proposed method was tested using an eight-element uniform circular array. Since the small-size array is horizontally placed, it is incapable of providing precise elevation discrimination, and so, the evaluation focused on the arrival azimuths. The scenarios were simulated using the image source method [13] to control reverberation time. The reverberation time $T60$ is set to 200 milliseconds in the simulated experiment. The white noise was artificially added to the simulated signal at SNR of 10 dB. The continuous speech taken from the TIMIT [14] database was used as source signal. The signal was re-sampled to 8000 Hz.

The proposed method was compared with TF-CHB [4] and STMV [15]. TF-CHB is a typical sparsity-based method, in which the azimuths are estimated at individual bins and summarized across all bins. The STMV is a typical beamformer-based method, which steers the frequency-averaged covariance matrix to various directions. The directions with local maximum coherence are identified as the directions of speech sources. The TF-CHB and STMV determine the number of sources in a way similar to the proposed method and both methods perform hypothesis test at 1° intervals. All three methods employed 256-point DTFT and 32 milliseconds frames

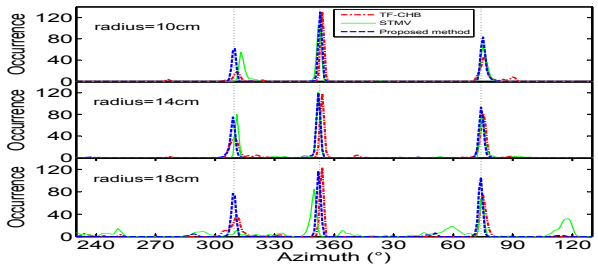

Fig. 4: Histograms of output azimuths under various array radius. The dotted lines denote the true azimuths.

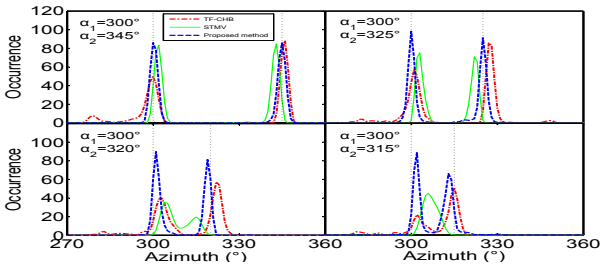

Fig. 5: Histograms of output azimuths with variant azimuth spaces of two speech sources. The dotted lines denote the true azimuths.

without frame overlap and conducted on the continuous speech segments with duration of 1.6 seconds.

The first experiment compared the influences of the array radius on performance. Three speakers were respectively located at a horizontal distance of 1.15 m, 0.86 m, and 0.63 m from the array center, and at the azimuth angles of 74°, 309.6°, 352.7°. The experimental setup is similar to AV16.3 corpus [16]. The array radius is respectively set to 10 cm, 14 cm, and 18 cm. The histograms of the output azimuths are plotted in Fig. 4. All three methods perform well on the array with 10 cm radius, where the spatial aliasing is not so serious. With the increasing of radius, the aliasing becomes more serious. On the microphone pair with 36 cm space, for example, there are at most eight time delay candidates that correspond to a given phase difference. However, the proposed method performs even better on the large-radius array than on the small-radius arrays. On contrast, STMV and TF-CHB are significantly deteriorated by the serious spatial aliasing. It should be mentioned that TF-CHB only utilizes the azimuth histogram instead of the time delay histogram, and so, it can not well treat spatial aliasing. This experimental results confirmed the superiority of the proposed method in spatial anti-aliasing.

The second simulated experiment investigated the spatial resolution of three methods. Two speakers were located at various azimuth spaces. Fig. 5 illustrates the azimuth histograms for different spacings of azimuth angles. The spatial resolution of TF-CHB is better than STMV because TF-CHB uses the azimuth histogram. STMV is incapable of distinguishing two sources when their azimuth spacing is less than 25°. The proposed method has the best spatial resolution among the three methods.

Table 1: Performance comparison on AV16.3 data set.

Algorithm	RMSE	PDR	FDR
TF-CHB	4.07°	74%	7.5%
STMV	3.07°	77%	8.1%
Proposed method	2.71°	83%	1.9%

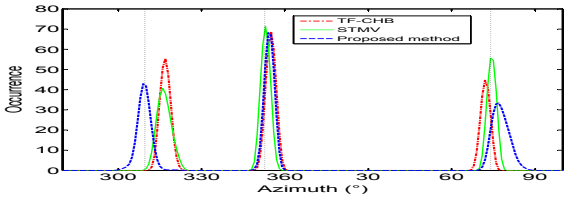

Fig. 6: Histogram of output azimuths for AV16.3 data set. The dotted lines denote the true azimuths.

Table 2: Computational load comparison.

Algorithm	Complex multiplication	Complex add
TF-CHB	16,180,000	2,323,000
STMV	373,700,000	373,740,000
Proposed method	345,000	1,280,000

The third experiment was conducted in real environment. The real data was taken from the publicly available AV16.3 corpus [16]. The signal used in this evaluation is the fourth fragment of the corpus recording, which is labeled "seq37-3p-0001". The signals were re-sampled to 8000 Hz. The radius of microphone array is 10 cm. The azimuth histogram is plotted in Fig. 6. The detected sources are separated into two categories, namely the correctly detected sources and the incorrectly detected sources. The detection is considered to be correct if the estimated azimuth deviates no more than 8° from the actual azimuth of any source. The incorrectly detected sources consist of the ghost sources (detected but non-existing sources) and the inaccurately detected sources. In this experiment, the incorrectly detected sources are seldom present, and so, RMSE can be utilized to evaluate the absolute error between the actual azimuths and the estimated azimuths. Besides, the positive detection rate (PDR) (i.e., the ratio of the number of correctly detected sources to the total number of sources) and the false detection rate (FDE) (i.e., the ratio of the number of incorrectly detected sources to the total number of sources) are used to evaluate the detection correctness. The RMSE, PDR and FDR are summarized in Table 1. The experimental result shows that the proposed method outperforms TF-CHB and STMV.

At last, the computational load of three methods are compared. The numbers of complex multiplication and complex add used by three methods are summarized in Table 2. The result shows that the computational load of the proposed method is much smaller than TF-CHB and STMV. The STMV has a heavy computational load because of the two-dimensional (360×90) grid search. Both the numbers of complex multiplication and complex add used by STMV are more than 250 times than the proposed method. The number of complex multiplication used by TF-CHB is nearly 4.7 times as much as the proposed method and the number of complex add is approximately 1.8 times than the proposed method. The proposed method outperforms other two methods in computational efficiency.

6. CONCLUSIONS

This paper proposes a sparsity-based method to localize multiple speech sources by utilizing time delay histograms on a planar array. Because the time delay histogram has the intrinsic advantages in spatial anti-aliasing and high spatial resolution, the proposed method can be applied on large-size arrays and can discriminate two closely located sources. Moreover, the proposed method exhibits high computational efficiency.

7. REFERENCES

[1] H. Krim and M. Viberg, "Two decades of array signal processing research: The parametric approach," *IEEE Signal Process. Mag.*, vol. 13, pp. 67-C94, 1996.

[2] S. Araki, H. Sawada, R. Mukai and S. Makino, "DOA estimation for multiple sparse sources with normalized observation vector clustering," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Toulouse, France, 2006, pp. 33C-36.

[3] Z. Wenyi and D. Rao, "A two microphone-based approach for source localization of multiple speech sources," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 18, no. 8, pp. 1913–1928, 2010.

[4] A. Torres, M. Cobos, B. Pueo, and J. Lopez, "Robust acoustic source localization based on modal beamforming and time-frequency processing using circular microphone arrays," *J. Acoust. Soc. Am.*, vol. 132, no. 3, pp. 1511C-1520, 2012.

[5] Y. X. Zou, W. Shi, B. Li, et al, "Multisource DOA estimation based on time-frequency sparsity and joint inter-sensor data ratio with single acoustic vector sensor," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4011–4015, May 2013.

[6] M. Ren and Y. Zou, "A novel multiple sparse source localization using triangular pyramid microphone array," *IEEE Signal Process. lett.*, vol. 19, no. 2, pp. 83C-86, 2012.

[7] M. Kuhne, R. Togneri, and S. Nordholm, "Robust source localization in reverberant environments based on weighted fuzzy clustering," *IEEE Signal Process. lett.*, vol. 16, no. 2, pp. 85C-88, 2009.

[8] J. Chen, J. Benesty, and Y. Huang, "Time Delay Estimation in Room Acoustic Environments: An Overview," *EURASIP J. on App. Signal Process*, pp. 1C-19, 2006.

[9] C. Liu, B. Wheeler, W. OBrien, R. Bilger, C. Lansing, and A. Feng, "Localization of multiple sound sources with two microphones," *J. Acoust. Soc. Am.*, vol. 108, no. 4, pp. 1888C-1905, 2000.

[10] D. Ying, F. Li, Z. Huang, G. Zhan, Y. Yan, "A closed-form method of spatial de-aliasing for multiple speech source localization," *Global. Sip*, 2015.

[11] Yilmaz, O. and Rickard, S., "Blind separation of speech mixture via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, 2004.

[12] D. Ying and Y. Yan, "Robust and fast localization of single speech source using a planar array," *IEEE Signal Process. lett.*, vol. 20, no. 9, pp. 909–912, 2013.

[13] J. Allen and D. Berkley, "Image method for efficiency simulating smallroom acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943C-950, 1979.

[14] J. S. Garofolo, "Getting started with the DARPA TIMIT CDROM: An acoustic phonetic continuous speech database," in *Nat. Inst. Stand. Technol. (NIST)*, Gaithersburg, MD, USA, Dec. 1988.

[15] J. Krolik and D. Swingler, "Multiple broad-band source location using steered covariance matrices," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no.10, pp. 1481C-1494, 1989.

[16] G. Lathoud, J. Odobez, and D. Gatica-Perez, "AV16.3: An audiovisual corpus for speaker localization and tracking," in *Proceedings of the 1st International Workshop on Machine Learning for Multimodal Interaction*, Martigny, Switzerland, 2004, pp. 192C-195.