Robust Multiple Speech Source Localization Based on Phase Difference Regression

Zhaoqiong Huang, Ge Zhan, Dongwen Ying, Ruohua Zhou, Jielin Pan and Yonghong Yan

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

huangzhaoqiong@hccl.ioa.ac.cn

Abstract

Spatial aliasing is a challenging issue that faced by most multiple speech source localization methods. Small-size arrays are widely used to avoid or mitigate spatial aliasing. But they deteriorate the coherence in low frequencies and degrade the performance of localization. This paper proposes a phase difference regression method for multiple speech source localization on a planar array. The time delay histogram is firstly applied to classify the frequency bins into clusters that correspond to speech sources, and then, the phase difference regression is conducted on each cluster. Since the error of the phase difference is limited in the range of $[-\pi, \pi]$, the proposed method avoids the ambiguity in the period number of phase. Although conventional regression method considers the period number, it does not bring significant advantage over the proposed method. The experimental results confirm the superiority of the proposed method on large-size arrays.

Index Terms: Speech source localization, phase difference regression, time delay histogram, spatial aliasing.

1. Introduction

Multiple speech source localization is widely used in numerous applications such as speech enhancement, speech separation, and speech recognition [1]. A large number of methods are proposed to localize multiple speech sources based on the property of speech sparse distribution in the time-frequency (TF) domain. Based on this property, the multiple-source signal model can be simplified to the single-source model, so the Direction-of-arrival (DOA) of the bin-dominant source is easily derived from the simplified model at each bin. Many sparsity-based methods were presented in the past several decades [2] - [9].

In addition to the acoustic robustness, the spatial aliasing is a challenging issues that are faced by the sparsity-based methods. Limiting the inter-microphone space is a simple way to avoid spatial aliasing [6], [10], whereas the small space will deteriorate the coherence at some low frequencies [11]. Traversing all candidates and selecting the most optimal candidates is another way to resolve spatial aliasing [3], [12]. Nevertheless, for a large-size array, there are many microphone pairs, and the candidates from different pairs will form numerous combinations. Traversing all possible combinations may lead to heavy computational load. A closed-form method of spatial de-aliasing for multiple speech source localization has been presented for real-time speech source localization [13]. But this method does not perform well under serious aliasing condition or adverse environment. This paper proposes a method using phase difference regression which is specially designed for the periodic variable, so the spatial aliasing is avoided just by limiting the phase difference error between the straightforward phase difference and the DOA-derived phase difference to $[-\pi, \pi]$. Although phase difference regression is widely utilized to localize speech sources [3], there are seldom methods reported to estimate DOA using planar arrays. The range of phase difference error is usually ignored in conventional methods.

The critical issue is to partition TF bins into several clusters, and then, the regression can be conducted on each cluster. Because the histogram analysis is a simple method to estimate DOAs with high spatial resolution and spatial anti-aliasing [14], the histogram analysis is used to determine the number of sources and estimate the initial DOAs. 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. These delays are combined to estimate the initial DOAs, which are chosen as the supervised information for bins classification. Eventually, the DOA of each source is estimated by means of regression over its associated phase differences.

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 [15]. Based on this property, the signal received by the kth microphone is represented in frequency domain as

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

where

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

where $0 \leq \omega_f \leq 2\pi$ 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. The phase difference of Fourier coefficients on the mth microphone pair, (k_1, k_2), is represented as

$$\begin{aligned} \psi_m(\omega_f) &= \mathcal{F}(\angle Y_{k_1}(\omega_f) - \angle Y_{k_2}(\omega_f), 2\pi) \\ &= \omega_f \tau_m^{(d)}(\omega_f) - 2\pi h_{m,f} + \xi(\omega_f) \end{aligned} \tag{3}$$

where $\angle(.)$ denotes the phase operation, $h_{m,f}$ is an integer, $\xi(\omega_f)$ is the perturbation caused by acoustic interferences, and operation $\mathcal{F}(X, T)$ is defined as

$$-T/2 \leq \mathcal{F}(X, T) = X + T \times n \leq T/2, \tag{4}$$

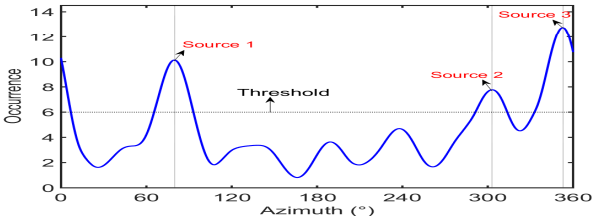

Figure 1: Azimuth histogram of three speech sources. The dotted lines denote the initial azimuths.

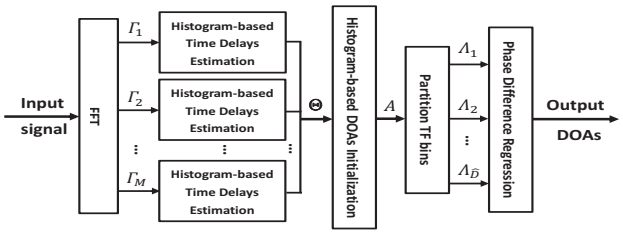

Figure 2: Block diagram of the proposed method.

where the integer n enables $\mathcal{F}(X,T)$ to be limited in the range of $[-T/2, T/2]$. At speech-dominant bins, the effect of noise signal can be disregarded, and then the time delay for the dominant source can be estimated from the phase difference,

$$\tau_m^{(d)}(\omega_f) \approx (\psi_m(\omega_f) + 2\pi h_{m,f})/\omega_f. \tag{5}$$

It should be mentioned that the subscript $(.)^d$ and the time index for the frequency bin are omitted in the followings.

For widely spaced array, there are several delay candidates at each frequency, the potential time delays are given by a set:

$$\begin{aligned} B_{m,f} = \Big\{ \tau \Big| \tau \approx (\psi_m(\omega_f) + 2\pi h_{m,f})/\omega_f, \\ -r_m/c \le \tau \le r_m/c \Big\}, \end{aligned} \tag{6}$$

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 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

$$\Gamma_m = \Big\{ B_{m,2}, \cdots, B_{m,F} \Big\}, \tag{7}$$

where the first frequency is disregarded since it does not contain the information of time delay. The time delay histogram is constructed on set $\Big\{ \Gamma_1, \ldots, \Gamma_M \Big\}$ to give the initial DOAs.

3. Proposed method

The basic idea of the proposed method is to partition TF bins into several clusters, and then, the DOA of each source is estimated by regression over its associated phase differences. Time delay histograms are firstly used to determine the number of sources and estimate the initial DOAs.

3.1. Initial estimation

By constructing the time delay histogram of the mth microphone pair using Γ_m, the impinging directions and the number of speech sources can be determined, where each significant peak with high occurrence is identified as a speech source. It is expressed as

$$\Big[\widehat{\tau}_{m,1}, \cdots, \widehat{\tau}_{m,I_m} \Big] = \mathcal{H}\Big(\bigcup_f B_{m,f} \Big), \tag{8}$$

where $\mathcal{H}(.)$ denotes the histogram operation, $\bigcup$ denotes the union operation and 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 time delays from all microphone pairs are expressed as a set

$$\Theta = \bigcup_m \Big\{ \widehat{\tau}_{m,1}, \cdots, \widehat{\tau}_{m,I_m} \Big\}. \tag{9}$$

In this paper, the DOA is represented by a unit direction vector γ, which can be derived from the azimuth α and elevation β of the source. For a planar array, every two delays (τ_1, τ_2) can determine a DOA, which is given by

$$\Big[\widehat{\alpha}_{\tau_1,\tau_2}, \widehat{\beta}_{\tau_1,\tau_2} \Big] = \mathcal{G}(\tau_1, \tau_2), \quad \tau_1 \ne \tau_2, \tag{10}$$

where $\mathcal{G}(\cdot)$ is a regression function that is determined by the array topology, the detail of which is given in reference [16]. Although we do not know to which source each time delay in Θ corresponds, we can make an hypothesis test to combine every two time delays. Three facts hold truth in these combinations. The first case is that the two delays are associated with the same speech source, where the determined azimuth is often distributed around the actual azimuth of this source. The second case is that the two delays belong to different sources, where the function in (10) may have no output or the outputs are randomly distributed. The third case is that the two delays correspond to the same microphone pair where the function in (10) has no output. All tested DOAs are expressed by a set,

$$A = \Big\{ (\alpha, \beta) \Big| [\alpha, \beta] = \mathcal{G}(\tau_1, \tau_2); \forall \tau_1, \tau_2 \in \Theta, \tau_1 \ne \tau_2 \Big\}. \tag{11}$$

The azimuths in set A are described by set $A^{(\alpha)}$. We construct a histogram to describe the distribution of azimuths. Based on these facts, each significant peak of the histogram corresponds to a speech source, and the number of sources is determined by counting the significant peaks, as shown in Fig. 1. The azimuth estimation using the histogram is represented as

$$\Big[\widehat{\alpha}_1, \ldots, \widehat{\alpha}_{\hat{D}} \Big] = \mathcal{H}\Big(A^{(\alpha)} \Big), \tag{12}$$

where $\hat{D}$ is the estimation of the source number. The elevation is determined based on the azimuth. The samples with an azimuth similar to the dth estimate is given by a set,

$$A_d^{(\beta)} = \Big\{ \beta \Big| (\alpha, \beta) \in A, |\alpha - \widehat{\alpha}_d| < \delta \Big\}, \tag{13}$$

where δ is empirically determined. The dth estimate of the elevation is given by

$$\widehat{\beta}_d = \frac{1}{|A_d^{(\beta)}|} \sum_{\beta \in A_d^{(\beta)}} \beta, \quad d = 1, \cdots, \hat{D}. \tag{14}$$

Finally, the initial estimate of the dth source is expressed as

$$\widehat{\gamma}_d = \Big[\cos\widehat{\alpha}_d \cos\widehat{\beta}_d \;\; \sin\widehat{\alpha}_d \cos\widehat{\beta}_d \;\; \sin\widehat{\beta}_d \Big]^T. \tag{15}$$

3.2. Phase difference regression

The DOAs of speech sources are refined by phase difference regression. The TF bins are firstly partitioned to various clusters by using the initial DOAs as the supervised information and each cluster corresponds to a speech source. So the multiple source localization is simplified to single source localization. The key point is to classify all bins to each source. A distance from a given bin to the dth source is defined as

$$L(f,d)=\sum_{m=1}^{M}\Big|\mathcal{F}(\psi_m(\omega_f)-\omega_f r_m \mathbf{g}_m^T\hat{\boldsymbol{\gamma}}_d/c,2\pi)\Big|, \quad (16)$$

where the unit vector $\mathbf{g}_m=[g_{m,1},g_{m,2},0]^T$ denotes the direction of the mth microphone pair. Their third dimension being set to zero indicates that all microphones lie in a plane. By using the distance, each bin is classified to the dominant speech source. The classification is expressed as

$$\mathcal{I}'\Big(f\Big)=\arg\min_{d\in[1:\hat{D}]}L(f,d). \quad (17)$$

Accordingly, the bins correspond to the dth source is given by

$$\Lambda_d=\Big\{f\big|f\in[2:F],\quad \mathcal{I}'(f)=d\Big\}. \quad (18)$$

The unit direction vector of the dth source is estimated by regression over all phase differences in Λ_d. For a given TF bin, the phase difference is expressed as Eq. (3). For a given DOA, the phase difference is expressed as

$$\widehat{\psi}_m(\omega_f)=\omega_f r_m \mathbf{g}_m^T\boldsymbol{\gamma}_d/c. \quad (19)$$

The cost function is defined as the mean square error between the two phase differences, which is given by

$$\varepsilon(\boldsymbol{\gamma}_d)=\sum_m\sum_{f\in\Lambda_d}\Big|\mathcal{F}\big(\psi_m(\omega_f)-\widehat{\psi}_m(\omega_f),2\pi\big)\Big|^2, \quad (20)$$

The phase difference error is denoted as $\zeta_m(\omega_f)=\psi_m(\omega_f)-\widehat{\psi}_m(\omega_f)$. DOA is estimated by minimizing $\varepsilon(\boldsymbol{\gamma}_d)$, as following:

$$\begin{aligned}\widehat{\boldsymbol{\gamma}}_d&=\min_{\boldsymbol{\gamma}}\varepsilon(\boldsymbol{\gamma}),\\ subject\quad to:&\boldsymbol{\gamma}^T\boldsymbol{\gamma}=1.\end{aligned} \quad (21)$$

The optimal estimator in sense of (21) is constructed by using the Kuhn-Tucker necessary condition for constrained minimization. The gradient Lagrange equation is given by

$$Z(\boldsymbol{\gamma}_d,\mu)=\varepsilon(\boldsymbol{\gamma}_d)+\mu(\boldsymbol{\gamma}_d^T\boldsymbol{\gamma}_d-1), \quad (22)$$

where μ is the Lagrange Multiplier. When a group of integer n computed, Eq. (22) can be confirmed to be a concave function with only one minimum. From $\nabla\boldsymbol{\gamma}_d Z(\boldsymbol{\gamma_d},\mu)=0$, the closed-form solution to DOA is given by

$$\begin{pmatrix}\widehat{\gamma}_{1,d}\\ \widehat{\gamma}_{2,d}\end{pmatrix}=\Big[\sum_m\sum_{f\in\Gamma_d}\omega_f^2 r_m^2\mathbf{g}_m^{'}\mathbf{g}_m^{'T}/c\Big]^{-1}$$

$$\times\Bigg[\sum_m\sum_{f\in\Gamma_d}\Big(\mathcal{F}\Big(\zeta_m(\omega_f),2\pi\Big)+\omega_f r_m\mathbf{g}_m^T\hat{\boldsymbol{\gamma}}_d/c\Big)\omega_f r_m\mathbf{g}_m^{'}\Bigg],$$

$$\widehat{\gamma}_{3,d}=\sqrt{1-\widehat{\gamma}_{1,d}^2-\widehat{\gamma}_{2,d}^2}, \quad (23)$$

where $\mathbf{g}_m^{'}=[g_{m,1},g_{m,2}]^T$.

The error of the phase difference is limited in the range of $[-\pi,\pi]$, so the proposed method avoids the ambiguity in the period number of phase.

4. Implementation

The block diagram of the proposed method is shown in Fig. 2, where the histogram analysis has been used twice. One is to estimate time delays of microphone pairs, and the other is to give the initial azimuths in order to estimate the initial DOAs. 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

$$\triangle=O_{avg}+\eta(O_{max}-O_{avg}), \quad (24)$$

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 algorithm is summarized in Algorithm 1.

Algorithm 1 : DOAs estimation

1: Calculate phase differences at all frequencies using (3).
2: Calculate time delay candidates at all frequencies using (5), (6) and (7).
3: Construct the time delay histograms to estimate the pairwise delays using (8) and (9).
4: Calculate the DOAs of every two delays in Θ using (10).
5: Construct the azimuth histogram and determine the number of speech sources $\widehat{D}$.
6: Calculate the initial DOAs using (12) - (15).
7: Partition TF bins to each source using (16), (17) and (18).
8: Regression over all phase differences corresponding to each source to estimate the final DOAs using (23).

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. The evaluation focused on the arrival azimuths. The scenarios were simulated using the image source method [17] to control reverberation time. The traffic noise was artificially added to the simulated signal at SNR of 10 dB. The continuous speech taken from the TIMIT [18] database was used as source signal. The signal was re-sampled to 8000 Hz.

The proposed method was compared with Circular Harmonics Beamforming (CHB) [4], Steered minimum variance (STMV) [19] and MUltiple SIgnal Classification (MUSIC). CHB is a typical sparsity-based method, in which the grid search is conducted to find the azimuths of the bin dominant sources. The speech sources are eventually identified by the azimuth histogram. The STMV is actually 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 MUSIC is a famous signal subspace method that tests the orthogonality between the noise subspace and the steering vector by grid search. The CHB and STMV determine the number of sources by counting the number of significant peaks in the histogram or the spatial spectrum power, just similar to the proposed method. Because MUSIC can not determine the number of sources, the number is assumed to be known in all following experiments. CHB, STMV, and MUSIC perform hypothesis grid search at 1° intervals.

The first experiment compared the influences of the array radius and reverberation on performance. Three speakers were

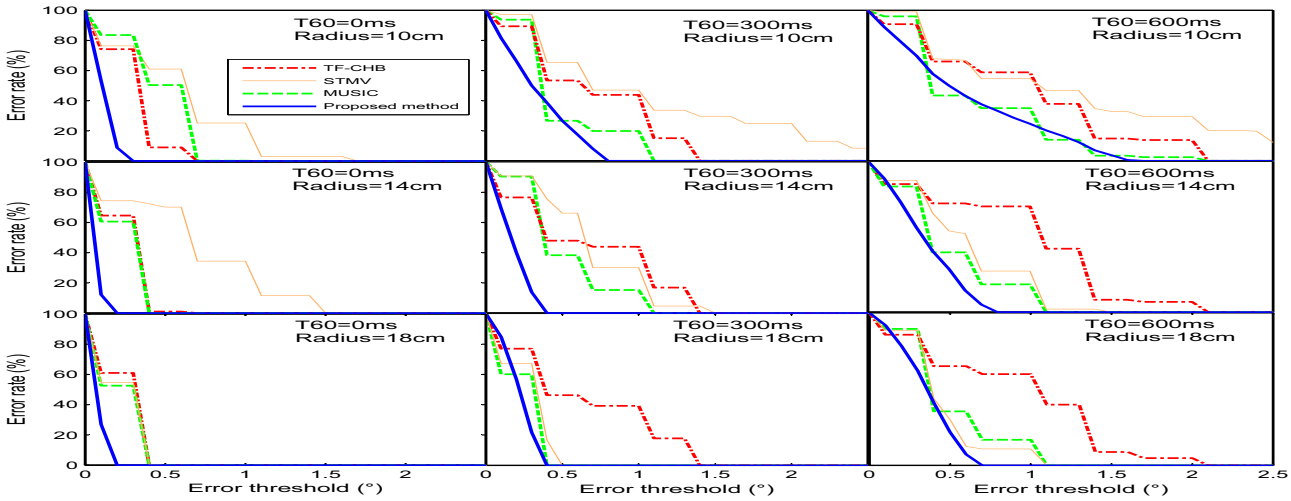

Figure 3: Error rate versus error threshold under various array radii and reverberations.

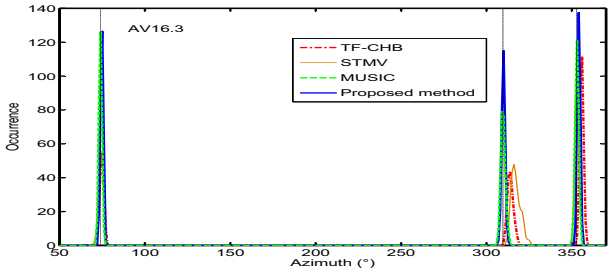

Figure 4: Histogram of output azimuths for AV16.3 data set. The dotted lines denote the true azimuths.

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

Algorithm	RMSE	PDR	FDR
TF-CHB	4.07°	74%	7.5%
STMV	3.00°	81%	12%
MUSIC	0.90°	92%	0%
Proposed method	1.01°	94%	0.1%

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 [20]. Experiments conducted on the continuous speech segments with duration of 1.6 seconds. The array radius is respectively set to 10 cm, 14 cm, and 18 cm. The reverberation time (T60) was respectively set as 0ms, 300 ms, and 600 ms. The error rate versus error threshold under all conditions are plotted in Fig. 3. Two-thirds of the detection azimuths are used here to guarantee there are no incorrectly detected sources, because the incorrect result may have a large influence on the error rate. Generally speaking, the accuracy on large-size arrays are better than small-size of all methods because the spatial resolution of large-size array is higher. Results show that the proposed method achieves the best detection accuracy under all conditions. This experimental results confirmed the superiority of the proposed method in spatial anti-aliasing and under reverberation condition.

The second experiment was conducted in real environment. The real data was taken from the publicly available AV16.3 corpus. 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. 4. 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 is 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 results show that the proposed method outperforms TF-CHB and STMV and is competitive with MUSIC.

6. Conclusions

This paper proposes a phase difference regression method to localize multiple speech sources. Because the error of phase difference is limited in the range of [-π, π], the period number of phase is no longer considered in the regression. The proposed method significantly simplifies regression, especially on large-size planar arrays. Since the ambiguity in the period number of phase difference is resolved by the histogram and the regression, the proposed method can be applied on any size planar arrays.

7. Acknowledgment

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).

8. References

[1] H. Krim and M. Viberg, "Two decades of array signal processing research: The parametric approach," *IEEE Signal Process. Mag.*, vol. 13, pp. 67–94, 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. 33–36.

[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] J. Dmochowski, J. Benesty, S. Affes, 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. 1511–1520, 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. 83–86, 2012.

[7] M. Swartling, N. Grbic, and I. Claesson, "Direction of arrival estimation for multiple speakers using time-frequency orthogonal signal separation, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2006, vol. 4, pp. 833–836.

[8] Adalbjornsson, S.I., Kronvall, T., Burgess, S., Astrom, K., Jakobsson, A. "Sparse Localization of Harmonic Audio Sources", in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, On page(s): 117–129 Volume: 24, Issue: 1, Jan. 2016.

[9] Xionghu Zhong, Hopgood, J.R. "A Time–Frequency Masking Based Random Finite Set Particle Filtering Method for Multiple Acoustic Source Detection and Tracking", in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, On page(s): 2356–2370, Volume: 23, Issue: 12, Dec. 2015.

[10] 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. 85–88, 2009.

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

[12] 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. 1888–1905, 2000.

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

[14] Z. Huang, G. Zhan, D. Ying, and Y. Yan, "Robust Multiple Speech Source Localization Using Time Delay Histogram," *ICASSP*, pp. 3191–3195, 2016.

[15] 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.

[16] 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.

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

[18] 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.

[19] 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. 1481–1494, 1989.

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