INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Multiple Sound Source Counting and Localization Based on Spatial Principal Eigenvector

Bing Yang, Hong Liu, Cheng Pang

Key Laboratory of Machine Perception,
Shenzhen Graduate School, Peking University, China

bingyang@sz.pku.edu.cn, hongliu@pku.edu.cn, chengpang@sz.pku.edu.cn

Abstract

Multiple sound source localization remains a challenging issue due to the interaction between sources. Although traditional approaches can locate multiple sources effectively, most of them require the number of sound sources as a priori knowledge. However, the number of sound sources is generally unknown in practical applications. To overcome this problem, a spatial principal eigenvector based approach is proposed to estimate the number and the direction of arrivals (DOAs) of multiple speech sources. Firstly, a time-frequency (TF) bin weighting scheme is utilized to select the TF bins dominated by single source. Then, for these selected bins, the spatial principal eigenvectors are extracted to construct a contribution function which is used to simultaneously estimate the number of sources and corresponding coarse DOAs. Finally, the coarse DOA estimations are refined by iteratively optimizing the assignment of selected TF bins to each source. Experimental results validate that the proposed approach yields favorable performance for multiple sound source counting and localization in the environment with different levels of noise and reverberation.

Index Terms: DOA estimation, source counting, spatial principal eigenvector, contribution removal, TF bin assignment.

1. Introduction

Multiple sound source localization is crucial in many acoustic signal processing tasks, such as speech dereverberation, noise reduction and blind source separation [1]. Several approaches have been proposed for multi-source localization, which can be broadly classified into three categories [2]: blind identification [3], angular spectrum [4, 5] and time-frequency (TF) processing [6]. However, most existing localization methods focus on improving accuracy and computational efficiency [7–9]. Other than these, an important issue in localization is source counting, since the number of sources is unknown in most practical applications and source localization will become more difficult without this information.

Subspace method, a kind of angular spectrum approach, has shown great effectiveness in source localization. A deep insight for eigen-analysis of parameterized spatial correlation matrix (PSCM) is given in [10]. It shows that when the PSCM is steered toward the source location, the major eigenvalue is maximized while all other eigenvalues are minimized. The multichannel cross-correlation coefficient (MCCC) approach [11] and the subband weighting based method [12] estimate the location by reaching the minimum of the determinant of PSCM, which is essentially equivalent to minimizing the minor eigenvalues [10]. Some other subspace methods perform the eigendecomposition of the spatial correlation matrix [13,14]. The extracted principal eigenvector and other minor eigenvectors contain different subspace information. The orthogonality of the minor eigenvectors to the steering vector [13] is considered in the the multiple signal classification (MUSIC) approach [15]. With the direction information of dominant source, the principal eigenvector is highly correlated with the steering vector, which is considered in the expectation-maximization eigenvector clustering method [14]. Although this method can localize closely spaced speakers and detect multiple sources, it requires the number of source as a prior information. Recently, TF methods have attracted broad attention due to their source counting ability [2, 16–18]. Some of them solve source counting problem along the similar line to matching pursuit algorithm [19], in which the source with the largest contribution is found and then its contribution is removed iteratively until a stop criterion is satisfied [2, 18].

Motivated by these works, a spatial principal eigenvector based method is proposed to reliably estimate both the number and the DOAs of multiple speech sources. Firstly, single source dominant TF bins are selected to depress the overlapping effect caused by multiple sources and reverberation. Then, the spatial principal eigenvector is extracted to construct the contribution function, which takes full advantage of the high correlation between the principal eigenvector and the steering vector. After that, the number and the DOAs of sources are estimated by iteratively detecting the potential source with the largest contribution. Since the TF bins belonging to the detected source are removed, the interaction between sources is reduced, making it easy and accurate to find other sources. Finally, by optimizing the TF bin assignment, the TF bins dominated by the same source are gradually clustered and the refining DOA estimations are estimated from each cluster. The proposed method is able to achieve high counting success rate and DOA accuracy in the noisy and reverberant scenario.

2. Problem Formulation

Assume there are K far-field speech sources observed by an array of M microphones in a noisy and reverberant scenario. The propagation attenuation from sources to microphones is supposed to be same. The received signal can be modeled in the short-time Fourier transform (STFT) domain as [14]:

$$\begin{aligned} X_m(n,\omega) = &\sum_{k=1}^{K} \alpha S_k(n,\omega) e^{-j\omega\tau_{\theta_k,m}} \\ &+ \sum_{k=1}^{K} S_k(n,\omega) H_m(\omega,\theta_k) + V_m(n,\omega), \end{aligned} \tag{1}$$

where $m \in \{1, 2, \ldots, M\}$ is the microphone index, $k \in \{1, 2, \ldots, K\}$ is the source index, n and ω denote the time frame index and frequency point index, respectively. Here, α represents the propagation attenuation factor, θ_k denotes the DOA of k-th source, $\tau_{\theta_k,m}$ is the time of arrival from

http://dx.doi.org/10.21437/Interspeech.2017-940

the k-th source to the m-th microphone, $H_m(\omega, \theta_k)$ represents the STFT transformed impulse response of reverberation, $X_m(n, \omega)$, $S_k(n, \omega)$ and $V_m(n, \omega)$ denote the STFT of the received signals, the source signals and the additive ambient noise in frequency domain, respectively. The signal model in Formula (1) can be rewritten in a vector form:

$$\boldsymbol{x}(n, \omega) = \sum_{k=1}^{K} S_k(n, \omega)[\boldsymbol{e}(\omega, \theta_k) + \boldsymbol{h}(\omega, \theta_k)] + \boldsymbol{v}(n, \omega), \quad (2)$$

where

$$\begin{aligned}
\boldsymbol{x}(n, \omega) &= [X_1(n, \omega), X_2(n, \omega), \ldots, X_M(n, \omega)]^T, \\
\boldsymbol{v}(n, \omega) &= [V_1(n, \omega), V_2(n, \omega), \ldots, V_M(n, \omega)]^T, \\
\boldsymbol{h}(\omega, \theta_k) &= [H_1(\omega, \theta_k), H_2(\omega, \theta_k), \ldots, H_M(\omega, \theta_k)]^T, \\
\boldsymbol{e}(\omega, \theta_k) &= [\alpha e^{-j\omega\tau_{\theta_k,1}}, \alpha e^{-j\omega\tau_{\theta_k,2}}, \ldots, \alpha e^{-j\omega\tau_{\theta_k,M}}]^T.
\end{aligned}$$

Here, $\boldsymbol{e}(\omega, \theta_k)$ represents the steering vector from the DOA θ_k to the array, and $\tau_{\theta_k,m}$ can be calculated using the known distance between sources and sensors. Generally, the number of sources is unknown in practice, so the task is to simultaneously estimate the number K and the DOAs $\{\theta_k\}$ of sound sources.

3. Source Counting and DOA Estimation

3.1. Time-frequency bin weighting

In a multi-source environment, microphone signals are always a mixture of multiple sources. The W-disjoint orthogonal (W-DO) assumption [20] is usually utilized to simplify the observed mixture model. But actually there are many sources with strong overlapping in certain TF bins due to the influence of multiple sources and reverberation, and this assumption is rarely met in practice. Applying the WDO assumption to these TF bins will lead to a large modeling error, thus affecting the following decision. To depress the influence of overlapping, single source dominant TF bins are selected by a TF weighting scheme. The TF weight for the (n, f)-th bin is expressed as:

$$W_{\mathrm{TF}}(n, f) = \begin{cases} 1, & if\ \overline{r}(n, f) > r_{\mathrm{th}} \\ 0, & otherwise \end{cases}, \quad (3)$$

where f represents the frequency bin index, $\overline{r}(n, f)$ is the mean value for the correlation coefficients of all microphone pairs, and r_{th} denotes a predefined threshold for $\overline{r}(n, f)$. The correlation coefficient $r_{m,m'}(n, f)$, corresponding to the (m, m')-th microphone pair, is defined as [18]:

$$r_{m,m'}(n, f) = \frac{R_{m,m'}(n, f)}{\sqrt{R_{m,m}(n, f)}\sqrt{R_{m',m'}(n, f)}}, \quad (4)$$

where $R_{m,m'}(n, f)$ denotes the magnitude of the cross power spectrum, which is computed as:

$$R_{m,m'}(n, f) = \sum_{\omega \in \mathbb{S}_f} |X_m(n, \omega) X^*_{m'}(n, \omega)|, \quad (5)$$

where $\mathbb{S}_f$ is the set of adjacent frequency points belonging to the f-th frequency bin.

3.2. Source counting and coarse DOA estimation based on spatial principal eigenvector

For each single source dominant TF bin, the principal eigenvector contains the direction information of the dominant source and shows a high correlation with the steering vector [14]. Therefore, a spatial principal eigenvector based method is proposed to estimate the number and the coarse DOAs of speech sources, by matching the spatial principal eigenvectors with the steering vector from all candidate directions.

The spatial correlation matrix for the (n, f)-th TF bin is estimated from N_f frequency points:

$$\hat{\boldsymbol{R}}(n, f) = \frac{1}{N_f} \sum_{\omega \in \mathbb{S}_f} \boldsymbol{x}(n, \omega)\boldsymbol{x}^H(n, \omega), \quad (6)$$

where N_f is the number of frequency points in the f-th frequency bin. Then, the eigen-decomposition of $\hat{\boldsymbol{R}}(n, f)$ is performed to obtain the eigenvalues $\{\lambda_m(n, f)\}$ and the eigenvectors $\{\boldsymbol{q}_m(n, f)\}$, in which the largest eigenvalue and corresponding spatial principal eigenvector are denoted as $\lambda_1(n, f)$ and $\boldsymbol{q}_1(n, f)$, respectively. These eigenvalues, indicating the energy of sources and noise, are utilized to form a confidence weight measuring the dominance of single source. The confidence weight $W_{\mathrm{C}}(n, f)$ is computed as [17]:

$$W_{\mathrm{C}}(n, f) = \frac{\lambda_1(n, f)}{\frac{1}{M-1}\sum_{m=2}^{M} \lambda_m(n, f)}. \quad (7)$$

By adding the confidence weight to the correlation of the spatial eigenvector and the steering vector, a weighted similarity is formed to measure the local contribution in the (n, f)-th TF bin, for the source from the candidate direction θ_k:

$$S(n, f, \theta_k) = W_{\mathrm{C}}(n, f)^{\beta} |\boldsymbol{q}_1(n, f)^H \overline{\boldsymbol{e}}(f, \theta_k)|^2, \quad (8)$$

where $\theta_k \in \{0, 1, \ldots, 359\}$, β is an adjustable variable that controls the influence of the confidence weight, and $\overline{\boldsymbol{e}}(f, \theta_k)$ denotes the mean steering vector pointing to the DOA θ_k for f-th frequency bin

$$\overline{\boldsymbol{e}}(f, \theta_k) = \frac{1}{N_f} \sum_{\omega \in \mathbb{S}_f} \boldsymbol{e}(\omega, \theta_k). \quad (9)$$

In practice, the normalized version of $\boldsymbol{e}(\omega, \theta_k)$ is utilized.

For a specific TF bin, $S(n, f, \theta_k)$ tends to generate peaks at the real directions of sources, and the highest peak corresponds to the direction of the dominant source. In order to make full use of the available information, $S(n, f, \theta_k)$ are summed over all single source dominant TF bins to form a DOA-related function $\sum_{n,f} W_{\mathrm{TF}}(n, f) S(n, f, \theta_k)$. Ideally, this derived function can generate multiple peaks with the same number of sources. However, multiple sources and reverberation will disturb the ideal peak distribution, making it more difficult to count by directly searching peaks. To achieve a reliable source counting, an iterative contribution removal algorithm is proposed. At each iteration, the algorithm finds one potential source with the largest contribution, and then the TF bins belonging to this potential source are detected and removed. The contribution function $\delta(\theta_k)$ of the k-th source is defined to measure the overall contribution of the source from the candidate direction θ_k:

$$\delta(\theta_k) = \sum_{n,f} W_{\mathrm{TF}}(n, f) W_{\mathrm{R}}(n, f) S(n, f, \theta_k), \quad (10)$$

where $W_{\mathrm{R}}(n, f)$ denotes the remaining weight which is used to remove the TF bins belonging to the detected sources. Using this weight, the contribution of detected sources is removed and the contribution function shows clear peaks at the DOA of remaining sources, thus reducing the interaction between sources.

At k-th iteration, the DOA with the largest contribution is:

$$\dot{\theta}_k = \arg \max_{\theta_k \in \{0, \ldots, 359\}} \delta(\theta_k). \quad (11)$$

Predefine three thresholds: δ_{th}, N_{th}, $N_{\Delta\mathrm{th}}$. The number of the remaining TF bins is denoted by N_{rem}, and the difference

Algorithm 1: Source counting and DOA estimation

```
Input:   $\{\boldsymbol{x}(n,\omega)\}$, $\{\boldsymbol{e}(\omega,\theta_k)\}$
Output:  the estimated number of the sources $\hat{K}$, the
         DOA estimations $\{\hat{\theta}_k\}$
$k \leftarrow 0$        $\rightarrow$ Initialization of the source index
$\hat{K} \leftarrow 0$        $\rightarrow$ Initialization of the number of sources
calculate the TF weight $W_{\rm TF}(n,f)$ using (3)
compute the spatial correlation matrix $\hat{\boldsymbol{R}}(n,f)$ as (6)
take eigen-decomposition of $\hat{\boldsymbol{R}}(n,f)$, and calculate the
confidence weight $W_{\rm C}(n,f)$ as (7)
while true do
    $k \leftarrow k+1$
    compute the contribution function $\delta(\theta_k)$ as (10)
    find the DOA with the largest contribution $\dot{\theta}_k$ using (11)
    if any stop criterion is satisfied then
        Break;
    end if
    $\hat{K} \leftarrow \hat{K}+1$, $\hat{\theta}_k \leftarrow \dot{\theta}_k$
    update the remaining weight $W_{\rm R}(n,f)$ by (12)
end while
while true do
    update the belonging weight $W_{\rm B}(n,f,k)$ as (14)
    compute the cost function $J(\theta_k)$ as (15)
    update the estimated DOA $\hat{\theta}_k$ by (16)
    if $\Delta J_{\rm all} < J_{\Delta{\rm th}}$ then
        Break;
    end if
end while
return $\hat{K}$, $\{\hat{\theta}_k\}$
```

of $N_{\rm rem}$ between current iteration and the previous iteration is represented by $\Delta N_{\rm rem}$. The stop criterions of the iteration are described as follows: (a) $\delta(\dot{\theta}_k) < \delta_{\rm th}$; (b) $N_{\rm rem} < N_{\rm th}$ with $N_{\rm rem} = \sum_{n,f} W_{\rm TF}(n,f)W_{\rm R}(n,f)$; (c) $\Delta N_{\rm rem} < N_{\Delta{\rm th}}$. If none of the above-mentioned criterions is satisfied, the estimated number of the sources $\hat{K}$ is increased by one, and the estimated DOA for the k-th source $\hat{\theta}_k$ is updated as $\dot{\theta}_k$. The contribution of the source from the direction $\hat{\theta}_k$ is removed by updating $W_{\rm R}(n,f)$:

$$W_{\rm R}(n,f) = 0, \quad if \;\; S(n,f,\hat{\theta}_k) > S_{\rm th}, \tag{12}$$

where $S_{\rm th}$ is a predefined threshold for the similarity value.

After several iterations, the estimated number of the sources $\hat{K}$ and the coarse DOA estimations $\{\hat{\theta}_k\}$ are obtained, which serve as the initialization of the DOA refining procedure.

3.3. TF bin assignment for DOA refining

In the source counting process, the iterative TF bin removal will lead to a unbalanced number of TF bins utilized to localize each source, resulting in a set of coarse DOA estimations. To improve the accuracy of DOA estimation, an iterative refining algorithm is utilized to optimize the assignment of the TF bins to each source. At each iteration, the TF bins are reassigned to their dominant source, and then the DOAs are reestimated using these reassigned bins.

The dominant source for each TF bin is determined by the source with the largest contribution:

$$\hat{k}(n,f) = \arg \max_{k\in\{1,\ldots,\hat{K}\}} S(n,f,\hat{\theta}_k), \tag{13}$$

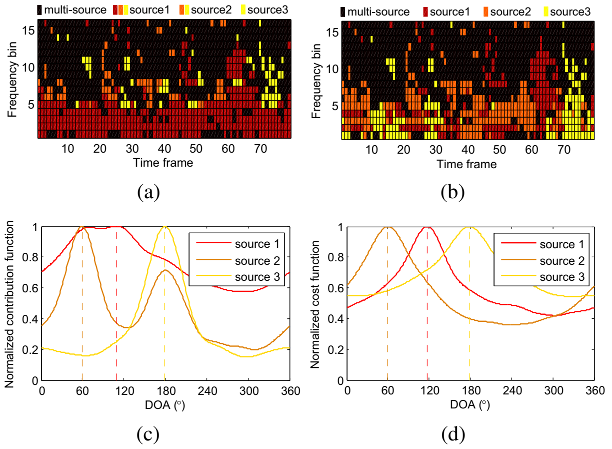


Figure 1: *A three-source instance in a reverberant and noisy room. (a) The TF bins used for counting each source. (b) The TF bins assigned to each source for refining the DOAs. (c) The normalized contribution function of each source. (d) The normalized cost function of each source.*

where $\hat{k}(n,f)$ is the dominant source index for the (n,f)-th bin. To assign the TF bins to their dominant source, the belonging weight $W_{\rm B}(n,f,k)$ is defined as:

$$W_{\rm B}(n,f,k) = \begin{cases} 1, & if \;\; k = \hat{k}(n,f) \\ 0, & otherwise \end{cases}. \tag{14}$$

Using this weight, the TF bins are clustered into multiple sets with the same number of sources. The cost function of the k-th source is defined to measure the contribution over all assigned TF bins, for the source from candidate DOAs:

$$J(\theta_k) = \sum_{n,f} W_{\rm TF}(n,f)W_{\rm B}(n,f,k)S(n,f,\theta_k). \tag{15}$$

The DOA of the k-th source is estimated from the TF bins assigned to this source, by finding the direction with the largest cost:

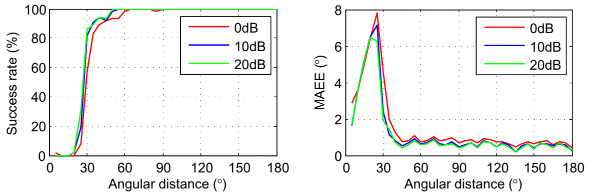
$$\hat{\theta}_k = \arg \max_{\theta_k\in\{0,\ldots,359\}} J(\theta_k). \tag{16}$$

The overall cost is $\sum_{k=1}^{\hat{K}} J(\hat{\theta}_k)$. The iteration will stop when the difference of the overall cost between current iteration and the previous iteration $\Delta J_{\rm all}$ is smaller than the threshold $J_{\Delta{\rm th}}$.

The details of the proposed algorithm for source counting and DOA estimation are described in Algorithm 1. A three-source instance is depicted in Fig. 1. Speech sources are located at 60°, 120° and 180°, respectively. The black TF bins labeled multi-source denote the TF bins dominated by multiple sources. It can be observed that the single source dominant T-F bins used for counting each source are not balanced and the first detected source tends to exploit more bins. After DOA refining, the number of TF bins assigned to each source is more reasonable. In addition, when compared with the contribution functions, the cost functions show relatively clearer peaks and the DOAs maximized the cost functions are more accurate.

4. Experiments and Discussions

The proposed method is evaluated in a reverberant room, which is characterized by the reverberation time $T_{60} = 250$ms. An 8-channel uniform circular array with a radius of 10cm is placed in the center of the room (4m × 4m × 3m). To create multi-source scenarios, sources are randomly located around the array with 1.5m distance, from 0° to 360° with an interval of



(a) (b)

Figure 2: *The performance versus different angular distances in a two-source environment. (a) The success rate of source counting. (b) The MAEE of DOA estimation.*

5°. Speech recordings from TIMIT database [21] are utilized as the source signals and the sampling rate is 16kHz. The image method [22] is used to generate the room impulse responds (RIRs) from sources to the sensors. The sensor signals are synthesized by convolving the clean speech signals with the generated RIRs. To control the signal-to-noise ratio (SNR), white Gaussian noise is properly scaled and added to each microphone signal. In the following processing, the captured array signals are set to 2s long. The frame length is 800 with 50% overlap and 1600 STFT coefficients are derived for each frame. The highest frequency of interest is set to 4kHz. The valid frequency points are divided into 16 frequency bins. The accuracy of source counting is measured by the success rate, that is the percentage of correct counting. The accuracy of DOA estimation is evaluated using the mean absolute estimated error (MAEE):

$$\mathrm{MAEE} = \frac{1}{N_{\mathrm{ins}} K_{\min}} \sum_{i=1}^{N_{\mathrm{ins}}} \sum_{k=1}^{K_{\min}} |\hat{\theta}_{k,i} - \theta_{k,i}|, \qquad (17)$$

where N_{ins} denotes the number of simulation instances, $K_{\min}$ = $\min(\hat{K}, K)$, and $\hat{\theta}_{k,i}$ and $\theta_{k,i}$ represent the estimated DOA and the true DOA, respectively. Here, N_{ins} is set to 200.

The spatial resolution of the proposed method is investigated. Fig. 2 shows the performance versus different angular distances from 5° to 180° with an interval of 5°, in a two-source environment with different SNRs. It can be observed that the proposed method hardly get the right number of sources with an angular distance smaller than 20°, because the interaction of close located sources will result in identifying them as one source. As the angular separation increases from 20° to 35°, both the counting error and the DOA error drop sharply. The performance remains good when the angular distance rises from 35° to 180°. Since the high correlation between the spatial principal eigenvector and the steering vector is taken into account, the proposed method can separate sources well for most cases.

Table I: *Comparison of the success rate for different methods*

SNR	methods	K=2	K=3	K=4	K=5
0dB	DC	77.00%	63.00%	42.50%	18.00%
	proposed	**92.50%**	**89.00%**	**86.00%**	**80.50%**
10dB	DC	86.50%	64.00%	46.50%	18.50%
	proposed	**95.00%**	**93.50%**	**91.00%**	**85.00%**
20dB	DC	87.00%	66.50%	47.00%	18.50%
	proposed	**96.50%**	**94.00%**	**93.00%**	**86.50%**

To evaluate the performance of source counting, the proposed source counting method is compared with the direct peak counting approach (labeled as DC). The DC approach estimates the number of sources from the contribution function by finding maximum peaks that larger than a predefined threshold. The smallest angular separation between sound sources is set to 35°. Table I depicts the comparison of success rate for different methods in different conditions. It is observed that our method performs better than the DC approach in all conditions. Since our approach removes the detected sources with largest contribution at each iteration, the interaction between sources is reduced. When the number of sources increases, the success rate of both methods degrade due to the increasing ambiguity caused by multi-source interaction. The DC approach is more sensible to the number of sources, e.g., when the number of sources is set to 5, the success rate of DC approach is relatively small while our method remains high success rate.

Table II: *Comparison of the MAEE (°) for different methods*

SNR	methods	K=2	K=3	K=4	K=5
0dB	EM	3.29	12.46	21.01	29.82
	SC	3.25	3.11	3.32	3.91
	proposed	**1.19**	**1.22**	**1.26**	**1.37**
10dB	EM	2.52	11.29	18.72	29.50
	SC	2.39	2.59	3.00	3.43
	proposed	**0.94**	**1.07**	**1.09**	**1.24**
20dB	EM	2.24	11.90	19.15	28.58
	SC	2.18	2.45	2.88	3.69
	proposed	**0.84**	**1.07**	**1.07**	**1.41**

The performance of DOA estimation is assessed by comparing three different methods: the expectation-maximization eigenvector clustering approach proposed in [14] (labeled as EM), the proposed source counting method with coarse DOA estimations (labeled as SC), and the proposed method with refined DOAs (labeled as proposed). For comparison, the EM approach is assumed with known number of sources, and the other two approaches are evaluated using the correctly counted instances. The minimum angular distance between sources is 35°. Table II illustrates the comparison of MAEE for these methods. It can be seen that the proposed method obtains a much lower MAEE after DOA refining. This confirms that the refining algorithm can correct the wrong association between the sources and TF bins according to the reestimated DOAs, thus improving the DOA accuracy. The proposed method also performs better than the EM approach. All the three approaches tend to have a worse performance when the number of sources or the noise level is relatively larger.

5. Conclusions

This paper proposes a spatial principal eigenvector based method for estimating both the number and the DOAs of multiple speech sources. By taking full advantage of the information contained in the spatial principal eigenvector, a weighed similarity is formed to evaluate the contribution of the source from candidate directions. The principal eigenvector based contribution removal approach can correctly count closely located sources due to the reduced interaction between sources. The DOA refining process optimizes the TF bins assignment to each source, thus improving the accuracy of DOA estimation. Experimental results show that the proposed method can count and localize multiple speech sources with high accuracy.

6. Acknowledgements

This work is supported by National High Level Talent Special Support Program, National Natural Science Foundation of China (No. 61340046, 61673030, U1613209), Specialized Research Fund for the Doctoral Program of Higher Education (No. 20130001110011), Natural Science Foundation of Guangdong Province (No. 2015A030311034). Specialized Research Fund for the strategic and prospective industrial development of Shenzhen city (No. ZLZBCXLJZI20160729020003).

7. References

[1] A. Alexandridis and A. Mouchtaris, "Multiple sound source location estimation and counting in a wireless acoustic sensor network," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 1–5, 2015.

[2] L. Wang, T. K. Hon, J. D. Reiss, and A. Cavallaro, "An iterative approach to source counting and localization using two distant microphones," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 6, pp. 1079–1093, 2016.

[3] A. Lombard, Y. Zheng, H. Buchner, and W. Kellermann, "TDOA estimation for multiple sound sources in noisy and reverberant environments using broadband independent component analysis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1490–1503, 2011.

[4] H. Do and H. F. Silverman, "SRP-PHAT methods of locating simultaneous multiple talkers using a frame of microphone array data," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 125–128, 2010.

[5] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," *Signal Processing*, vol. 92, no. 8, pp. 1950–1960, 2012.

[6] T. N. T. Nguyen, S. Zhao, and D. L. Jones, "Robust DOA estimation of multiple speech sources," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2287–2291, 2014.

[7] C. Pang, J. Zhang, and H. Liu, "Direction of arrival estimation based on reverberation weighting and noise error estimator," *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 3436–3440, 2015.

[8] D. Yook, T. Lee, and Y. Cho, "Fast sound source localization using two-level search space clustering," *IEEE Transactions on Cybernetics*, vol. 46, no. 1, pp. 20–26, 2016.

[9] J. Zhang and H. Liu, "Robust acoustic localization via time-delay compensation and interaural matching filter," *IEEE Transactions on Signal Processing*, vol. 63, no. 18, pp. 4771–4783, 2015.

[10] M. Souden, J. Benesty, and S. Affeso, "Broadband source localization from an eigenanalysis perspective," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1575–1587, 2010.

[11] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: an overview," *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. 1–19, 2006.

[12] W. Xue and W. Liu, "Direction of arrival estimation based on subband weighting for noisy conditions," *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 142–145, 2012.

[13] J. Benesty, J. Chen, and Y. Huang, "Microphone array signal processing," *Springer*, 2008.

[14] X. Xiao, S. Zhao, T. N. T. Nguyen, D. L. Jones, E. S. Chng, and H. Li, "An expectation-maximization eigenvector clustering approach to direction of arrival estimation of multiple speech sources," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6330–6334, 2016.

[15] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[16] Z. E. Chami, A. Guerin, A. Pham, and C. Serviere, "A phase-based dual microphone method to count and locate audio sources in reverberant rooms," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 209–212, 2009.

[17] S. Arberet, R. Gribonval, and F. Bimbot, "A robust method to count and locate audio sources in a multichannel underdetermined mixture," *IEEE Transactions on Signal Processing*, vol. 58, no. 1, pp. 121–133, 2010.

[18] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2193–2206, 2013.

[19] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[20] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[21] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium*, 1993.

[22] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.