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

Time Delay Histogram Based Speech Source Separation Using a Planar Array

Zhaoqiong Huang[1,2], *Zhanzhong Cao*[1,2], *Dongwen Ying*[1,2], *Jielin Pan*[1,2], *and Yonghong Yan*[1,2,3]

[1] Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences
[2] University of Chinese Academy of Sciences
[3] Xinjiang Laboratory of Minority Speech and Language Information Processing, Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences

{huangzhaoqiong,caozhanzhong,yingdongwen,panjielin,yanyonghong}@hccl.ioa.ac.cn

Abstract

Bin-wise time delay is a valuable clue to form the time-frequency (TF) mask for speech source separation on the two-microphone array. On widely spaces microphones, however, the time delay estimation suffers from spatial aliasing. Although histogram is a simple and effective method to tackle the problem of spatial aliasing, it can not be directly applied on planar arrays. This paper proposes a histogram-based method to separate multiple speech sources on the arbitrary-size planar array, where the spatial aliasing is resisted. Time delay histogram is firstly utilized to estimate the delays of multiple sources on each microphone pair. The estimated delays on all pairs are then incorporated into an azimuth histogram by means of the pairwise combination test. From the azimuth histogram, the direction-of-arrivals (DOAs) and the number of sources are obtained. Eventually, the TF mask is determined based on the estimated DOAs. Some experiments were conducted under various conditions, confirming the superiority of the proposed method.

Index Terms: Speech source separation, time delay histogram, direction-of-arrivals, planar array.

1. Introduction

Speech source separation using microphone arrays has received growing interest in numerous applications such as video conferencing, automatic speech recognition and human-computer interaction [1, 2]. The acoustic interference such as reverberation or environmental noise is the major challenging issue that faced by most speech source separation methods. Numerous methods are presented in the past several decades.

Independent component analysis (ICA) is widely used to separate multiple speech sources based on the assumption that the speech sources are statistically independent with each other [3] – [9]. The separation filter is determined by maximizing the interdependency between sources. Most ICA-based methods usually face two problems. One is the permutation ambiguity, and the other is these methods cannot determine the number of sources. Independent vector analysis (IVA) [8, 9] is a special realization of ICA, which avoids permutation ambiguity by applying the IVA on the full band, instead of an individual frequency. But it still requires the number of sources to be pre-estimated.

Spatial information of sources is another valuable clue for speech source separation, which is highly correlated with the inter-microphone time delays [10] – [21]. The time delay histogram is a simple and effective method that not only estimates delays, but also resist spatial aliasing [21]. The histogram-based methods are motivated by a fact that each frequency bin is dominated by a unique dominant source. This property refers to W-disjoint orthogonality [22]. Based on this property, the speech sources are identified by using the histogram of bin-wise time-delays. For two microphones, the number of sources and the time delays of each source are determined by counting the significant peaks in the histogram. Eventually, the time frequency (TF) masks can be estimated from the delays. At the state of the art, the histogram-based methods are generally applied on linear arrays [22] – [24], but seldom reported to be applied on planar array. However, the planar array has the superiority over the linear array in spatially discriminating multiple sources, which enables a good performance in speech source separation.

This paper proposes a method to separate multiple speech sources using the time delay histograms on the planar array. The time delays are firstly extracted from the histogram of each microphone pair by the peak picking method. The estimated delays on all pairs are then incorporated into an azimuth histogram by means of the pairwise combination test. The speech sources are identified from the histogram, where the number and direction-of-arrivals (DOAs) of speech sources are determined. By using the DOAs as the supervised information, the TF mask of each source is generated, and the permutation ambiguity problem [18] is resolved.

2. Signal model

Let's consider D speech sources that impinge on a K-element array in a far-field scenario. This scenario assumes that the size of the array aperture is small relative to the distance from each source to the array center, and so, the attenuation effect for signal propagation is disregarded. The relationship between the source signal and the recorded signal is described by the widely used free-field model [10]. The mixture received by the kth microphone in frequency domain is described as

$$Y_k(\omega_f) = \sum_{d=1}^{D} e^{-j\omega_f \varphi_{k,d}} S_d(\omega_f) + N_k(\omega_f), \tag{1}$$

where S_d denotes the signal emitted from the dth source, $\varphi_{k,d}$ denotes the propagation time from the dth source to the kth microphone, $j = \sqrt{-1}$ denotes the imaginary unit, f denotes the frequency index, ω_f denotes the angular frequency and $N_k(\omega_f)$ denotes the acoustic interference at the kth microphone.

Based on the speech sparsity property [22], Eq. (1) is simplified as

$$Y_k(\omega_f) = e^{-j\omega_f \varphi_{k,d}} S_d(\omega_f), \tag{2}$$

where d denotes the index of the dominated source. The environmental noise is taken as a nondirectional source, and noise-dominant bins are not taken into account. Therefore, the noise item is disregarded in the simplified model.

http://dx.doi.org/10.21437/Interspeech.2017-55

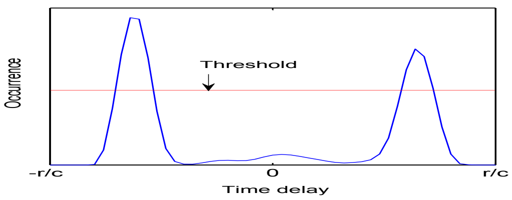

Figure 1: *Time delay histogram of two speech sources.*

For the K-element planar array, there are $M = K(K-1)/2$ pairs of microphones. On each pair, the time delay can be calculated from the phase difference of Fourier coefficients. There is an ambiguity in the period number of phase difference. The candidate of phase difference with the smallest absolute value is referred to as the minimal phase difference. At the fth bin of the mth microphone pair, (k_1, k_2), it is expressed as

$$\phi_{m,f} = \mathcal{F}\Big(\angle Y_{k_2}(\omega_f) - \angle Y_{k_1}(\omega_f), 2\pi\Big), \qquad (3)$$

where $\angle(.)$ denotes the operation of taking phase, $\mathcal{F}(X,T)$ is the function that regulates the periodical variable X into the range of $[-T/2, T/2]$, T denotes the period of the variable X, and thus $-\pi \le \phi_{m,f} \le \pi$. Accordingly, the time delay of the dominant source is given by

$$\widehat{\tau}_m^{(d)}(\omega_f) = (\phi_{m,f} + 2p_{m,f}\pi)/\omega_f, \qquad (4)$$

where $p_{m,f}$ denotes the number of periods. Because the time delay is constrained by the inter-microphone distance, the candidates for a time delay are given by the set

$$\begin{aligned} B_{m,f} = \Big\{\tau \Big| \tau = (\phi_{m,f} + 2p\pi)/\omega_f, \\ \frac{-r_m - c\phi_{m,f}}{cT_f} \le p \le \frac{r_m - c\phi_{m,f}}{cT_f}, T_f = 2\pi/\omega_f\Big\}, \end{aligned} \qquad (5)$$

where c denotes the sound velocity and r_m denotes the inter-microphone distance. It should be mentioned that the subscript $(.)^{(d)}$ and the time index for the TF bin are omitted for simplicity in the following. The TF bins, which are dominated by one source, have the same time delay. Therefore, the masks for speech source separation can be yielded by analyzing the bin-wise time delays.

3. Two microphone-based speech separation

Histogram is widely used to analyze the sources from the bin-wise time delays on two-microphone arrays, where the mth microphone pair is chosen as an example to introduce the speech separation with two microphones. The time delays of speech sources can be determined by applying histogram analysis on two microphones, where each significant peak is identified as a source. A time delay histogram for two speech sources is shown in Fig. 1, where the threshold is set by experience.

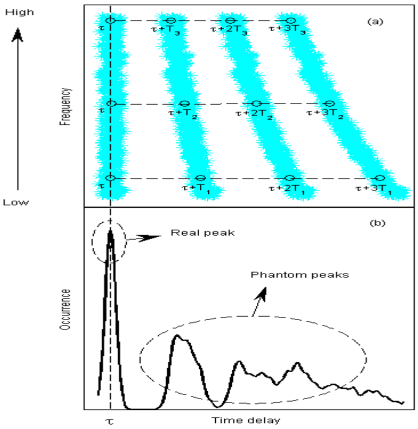

Figure 2: *Schematic illustration of the spatial anti-aliasing on the time delay histogram: (a) Scattered plots of bin-wise time delays; (b) Time delay histogram. τ is the real delay and the remaining delays ($\tau + pT_f$) are yielded by the aliasing error.*

The histogram of bin-wise time delays has a significant advantage in resisting spatial aliasing. When spatial aliasing occurs, all the candidates are utilized to construct the histogram. 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, which is the reason for the histogram to resist spatial aliasing. The capability of spatial anti-aliasing is schematically illustrated by Fig. 2. The histogram analysis can be expressed by a function,

$$\Big[\widehat{\tau}_{m,1}, \cdots, \widehat{\tau}_{m,I_m}\Big] = \mathcal{H}\Big(\bigcup_f B_{m,f}\Big), \qquad (6)$$

where I_m denotes the number of distinct peaks. In desirable acoustic conditions, I_m equals the source number D. Under adverse environments, however, I_m may be greater than or less than the number of real speech sources.

The TF mask of each source can be constructed using the estimated delays of two microphones. It relies on the sparseness property of source signals. Each TF bin can be classified to the active source. The dominant source at the fth bin is labeled as the source with the smallest distance, which is given by

$$\mathcal{D}_m\Big(f\Big) = \arg\min_{d\in[1:I_m]} \Big|\mathcal{F}(\phi_{m,f} - \omega_f\widehat{\tau}_{m,d}, 2\pi)\Big|. \qquad (7)$$

Accordingly, the dth source mask at the fth frequency is expressed as

$$\mathcal{M}_{m,d}(\omega_f) = \begin{cases} 1, \ if \ \mathcal{D}_m\Big(f\Big) == d \\ 0, \ otherwise. \end{cases} \qquad (8)$$

The dth separated speech signal received by the kth microphone is constructed by TF mask, which is represented as

$$X_{d,k}(\omega_f) = \mathcal{M}_{m,d}(\omega_f) Y_k(\omega_f). \qquad (9)$$

Therefore, by applying inverse short-term Fourier transform (STFT) to $X_{d,k}(\omega_f)$, the separated speech signal in time domain is obtained. The signal recorded by any microphone can be used to re-construct the signal of the dth source.

4. Planar array-based speech separation

However, there are two drawbacks of the two microphone-based separation method. The first is the poor capability in spatially discriminating multiple sources in the three-dimension space. The second is the unreliability in adverse environments. On the contrary, the planar array can utilize the redundant information provided by microphones to enhance the performance, and also discriminates sources in the three-dimension space. In this section, we introduce the planar array-based speech separation method, where the DOAs of speech sources are used as the supervised information to form source masks.

Table 1: *SIR (dB) comparison under various azimuth spacings.*

Method	(47°, 102°)		(50°, 95°)		(43°, 78°)		(50°, 75°)		(45°, 60°)	
	Mean	STD	Mean	STD	Mean	STD	Mean	STD	Mean	STD
IVA	8.80	3.33	8.73	3.24	8.11	2.88	7.74	2.88	**8.61**	2.12
CHB	7.22	1.75	7.33	1.87	4.98	1.97	5.06	**1.75**	4.91	**1.41**
TDH	**10.48**	**1.41**	**11.37**	**1.51**	**8.44**	**1.41**	**7.92**	1.81	7.07	1.66

4.1. DOAs estimation

The planar array consists of several microphone pairs, and the array topology is expressed by a group of vectors, $[\mathbf{g}_1, \cdots, \mathbf{g}_M]$, where $\mathbf{g}_m = [g_{m,1}, g_{m,2}, 0]^T$ denotes the unit directional vector between the locations of the mth pair of microphones. The time delay histogram of each microphone pair can yield several time delays as (6). The time delays from all microphone pairs are expressed as a set,

$$\Theta = \bigcup_m \left\{ \widehat{\tau}_{m,1}, \cdots, \widehat{\tau}_{m,I_m} \right\}. \tag{10}$$

An explicit DOA is determined by at least two time delays in Θ from unaligned microphone pairs. We conduct a pairwise combination test to combine every two delays to derive the DOA that can be expressed by an azimuth angle α and an elevation angle β, given by,

$$[\alpha, \beta] = \mathcal{G}\Big(\mathbf{g}^{(\tau_1)}, \mathbf{g}^{(\tau_2)}, \tau_1, \tau_2\Big), \tag{11}$$

where $\mathcal{G}(.)$ is a regression function, the details of which are given in [25]. All tested DOAs are expressed by a set,

$$A = \left\{ (\alpha, \beta) \Big| [\alpha, \beta] = \mathcal{G}\Big(\mathbf{g}^{(\tau_1)}, \mathbf{g}^{(\tau_2)}, \tau_1, \tau_2\Big); \forall \tau_1, \tau_2 \in \Theta \right\}. \tag{12}$$

In the combination test, only the DOAs of the correct combinations are likely to be distributed around the real DOAs. The incorrect combinations and the phantom delays are effectively suppressed by the test [21]. Since the planar array is placed with horizontal orientation, precise discrimination of the elevations is incapable of being provided, and the azimuth is the relatively reliable feature to discriminate different sources. We construct a histogram using the azimuths in the set A, which are denoted as the set $A^{(\alpha)}$. The number and DOAs of sources are determined by analyzing the significant peaks in the azimuth histogram. The estimates of azimuths are represented as

$$\Big[\widehat{\alpha}_1, \ldots, \widehat{\alpha}_{\widehat{D}}\Big] = \mathcal{H}\Big(A^{(\alpha)}\Big), \tag{13}$$

where $\widehat{D}$ is the estimate of the source number.

The elevation is determined based on the azimuth. The samples with an azimuth similar to the dth estimate are given by a set,

$$A_d^{(\beta)} = \left\{ \beta \Big| (\alpha, \beta) \in A, |\alpha - \widehat{\alpha}_d| < \delta \right\}, \tag{14}$$

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

$$\widehat{\beta}_d = \frac{1}{|A_d^{(\beta)}|} \sum_{\beta \in A_d^{(\beta)}} \beta, \quad d = 1, \cdots, \widehat{D}. \tag{15}$$

Finally, the DOA of the dth source is expressed as a unit directional vector as

$$\widehat{\gamma}_d = \Big[\cos\widehat{\alpha}_d \cos\widehat{\beta}_d \;\; \sin\widehat{\alpha}_d \cos\widehat{\beta}_d \;\; \sin\widehat{\beta}_d\Big]^T. \tag{16}$$

Table 2: *SIR (dB) comparison under various SNRs.*

Method	5 dB		10 dB		20 dB	
	Mean	STD	Mean	STD	Mean	STD
IVA	5.94	1.90	7.44	2.79	8.74	3.11
CHB	3.93	0.89	5.56	1.29	6.97	1.66
TDH	**6.31**	**0.51**	**8.49**	**0.85**	**10.20**	**1.28**

4.2. TF masks estimation

The spectrum is partitioned into several source masks, and each mask corresponds to a speech source, the DOA of which is selected as the supervised information. The distance from the fth bin to the dth source-labeled cluster is defined as the root mean square error between the minimal phase difference vector and the $\widehat{\gamma}_d$-derived phase difference vector [26], which is given by

$$L(f, d) = \sqrt{\frac{\sum_{m=1}^{M} \Big|\mathcal{F}(\phi_{m,f} - \omega_f r_m \mathbf{g}_m^T \widehat{\gamma}_d / c, 2\pi)\Big|^2}{M}}. \tag{17}$$

The dominant source at the fth bin is labeled as

$$\mathcal{D}\Big(f\Big) = \arg \min_{d \in [1:\widehat{D}]} L(f, d). \tag{18}$$

Accordingly, the dth source mask at the fth frequency is expressed as

$$\mathcal{M}_d(\omega_f) = \begin{cases} 1, \; if \;\; \mathcal{D}\Big(f\Big) == d \\ 0, \; otherwise. \end{cases} \tag{19}$$

Therefore, using (9) and inverse STFT, the separated speech signal in time domain is obtained. It should be mentioned that the permutation ambiguity faced by many frequency-domain blind source separation methods is resolved by using the DOAs as the supervised information. The algorithm of the proposed method is summarized in Algorithm 1.

Algorithm 1 : Speech source separation algorithm

```
1: Calculate time delay candidates using (3) - (5).
2: Estimate time delays of each microphone pair based on histogram analysis and obtain the delay sets using (6) and (10).
3: Determine the source number and estimate the DOAs of speech sources using (11) - (16).
4: Partition TF bins into clusters and obtain the mask of each source using (17) - (19).
5: Apply the inverse STFT to each separated signal as (9) and obtain the separated signal in time domain.
```

5. Evaluation

This section evaluates the proposed method by the simulated environment. The simulated scenarios were constructed using the image source method [27] to control the reverberation

Table 3: *SIR (dB) comparison under various reverberations.*

Method	200 ms		400 ms		600 ms	
	Mean	STD	Mean	STD	Mean	STD
IVA	**14.19**	4.56	8.07	2.93	5.99	2.35
CHB	9.54	2.37	6.59	1.62	4.51	**1.49**
TDH	14.14	**1.52**	**9.55**	**1.59**	**6.95**	1.50

time (RT60). The proposed method was tested using an eight-element uniform circular array and the radius is set to 10 cm. The continuous speech taken from TIMIT [28] database was used as source signal. There are 40 test samples for each condition in the experiment.

The IVA [8] and circular harmonic beamformer (CHB) [29] are used as the competing algorithm. IVA is a special realization of famous ICA method, wherein the permutation process is not required. CHB is a sound source localization method based on the bin-wise DOA estimation. The TF mask can be easily obtained by the bin-wise DOAs. CHB also plays a role of speech separation. TDH standing for "Time Delay Histogram" denotes the proposed method. Both TDH and CHB can determine the number of speech sources, while IVA cannot do it. For the sake of fairness, the number of speech sources is assumed to be known in advance. Moreover, the performance of separation can be assessed in terms of SIR (signal to interference ratio) without false and missed detections. All three methods employed 256-point STFT and 32 milliseconds frames with half-length overlap.

The first experiment investigates spatial resolution of three methods in a noise-free environment. Two speech sources are located at various spacings of azimuth. The reverberation time is set to 350 ms. The mean and standard deviation (STD) of SIR are averaged over all samples at each condition, as shown in Table 1. IVA discriminates the multiple sources based on the statistical independence between sources, whereas the remaining methods discriminate sources based on the spatial information that is contained in the phase of recorded signals. For the case of very small spacing (15°), the spatial difference is insignificant. IVA does not suffer from the small spacings of azimuth so much as the remaining methods, which enables TDH and CHB to be inferior to IVA. With general spacings of azimuth, however, TDH performs better than CHB and IVA. This experimental result is twofold. One is that the spatial difference is more reliable than the statistical independence. The other is that TDH more fully makes use of the spatial difference than CHB does.

The second experiment evaluates the performance under the room noise of the AV16.3 corpus [30]. The noise signal is artificially added to the reverberated source signal with various SNRs. The reverberation and azimuth spacing are respectively set to 350 ms and 55°. The mean and STD is summarized in Table 2. The third experiment evaluates the performance in a noise-free environment with various reverberations. The azimuth spacing is fixed to 55°. The experimental results are listed in Table 3. These two experiments evaluate the robustness of three methods over acoustic interferences. TDH generally performs best among the three methods. CHB and TDH are conducted based on the assumption of speech sparsity, which leads to the masking effect [22] on weak speech sources. Even in a very friendly environment, the speech sources cannot be completely separated because of the incomplete spectrogram that is provided by the TF masking. On contrast, IVA can accurately discriminate the speech sources if there are no acoustic interferences. Therefore, IVA may outperform TDH and CHB under interference-free conditions. For example, IVA performs slightly better than TDH under 200 ms in the third experiment.

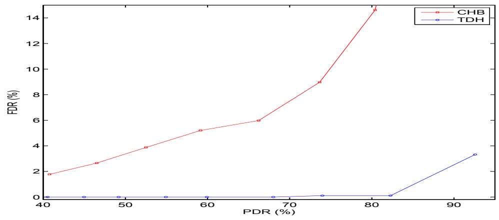

Figure 3: *ROC curve for counting the source number.*

IVA heavily relies on the assumption that the sources are statistically independent with each other, however, this assumption does not always hold truth. If the test sample does not well meet this assumption, the performance of separation will degrade a lot. On the contrary, CHB and TDH does not rely on this assumption, which enable a stationary performance. Therefore, the STD of IVA is larger than the remaining methods. In addition, CHB and TDH fully utilize the array topology to spatially discriminate speech sources whereas IVA does not. CHB estimates the DOAs in a bin-wise manner, and it is likely to suffer from acoustic interference such as environmental noise and reverberation. TDH estimates the DOAs from all bin-wise delays, instead of an individual bin-wise delay. Therefore, TDH can effectively resist reverberation and noise.

The last experiment compares the performance of counting the number of sources between CHB and TDH. The receiver operating performance characteristics (ROC) curve is used to access the performance, which can give a complete description of the relationship between PDR (positive detection rate) and FDR (false detection rate) by tuning the threshold in azimuth histogram. The array radius, RT60 and SNR are respectively set to 10 cm, 400 ms and 10 dB. There are totally 100×2 sources for the two-source evaluation and 100×3 sources for the three-source evaluation. For two-source scenario, all speech sources can be correctly detected without any false detection for TDH, i.e., PDR is 100% and FDR is 0%. While the PDR is 90% and FDR is 8% for CHB. The ROC curve for the three-source scenario is plotted in Fig. 3. The experimental results show that TDH outperforms CHB in counting the source number.

6. Conclusions

This paper proposes a novel method to separate multiple speech sources using a planar array. The histogram plays two roles. One is to extract the time delays of speech sources from the time delay histograms. The other is to summarize the DOAs of speech sources, which are estimated from the delays of speech sources from all microphone pairs. The experimental results showed that the proposed method has superiority not only in speech separation, but also in estimating the number of sources. The appealing property of the proposed algorithm is that it can separate multiple speech sources on arbitrary-size planar arrays, where the spatial aliasing does not affect on speech separation.

7. Acknowledgements

This study is supported by National Program on Key Basic Research Project (2013CB329302), National Natural Science Foundation of China (Nos. 61671442, 11590770) and the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant Nos. XDA06030100, XDA06030500).

8. References

[1] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu, W. Fabian, M. Espi, T. Higuchi, S. Araki, and T. Nakatani, “The NTT CHiME-3 system: advances in speech enhancement and recognition for mobile multi-microphone devices,” *Proc. Worksh. Automat. Speech Recognition Understanding*, 2015.

[2] K. Itakura, I. Nishimuta, Y. Bando, K. Itoyama, and K. Yoshii, ”Bayesian Integration of Sound Source Separation and Speech Recognition: A New Approach to Simultaneous Speech Recognition,” *Interspeech*, pp. 736–740, 2015.

[3] T. Lee, *Independent Component Analysis: Theory and Applications*. Boston, MA: Kluwer, 1998.

[4] T. Lee, M. Lewicki, M. Girolami, and T. Sejnowski, “Blind source separation of more sources than mixtures using overcomplete representations,” *IEEE Signal Processing Lett.*, vol. 6, pp. 87–90, 1999.

[5] A. Shirazi and B. Rao, “An ICA-SCT-PHD filter approach for tracking and separation of unknown time-varying number of sources,” *IEEE Trans. on Audio, Speech and Lang. Process.*, vol. 21, no. 4, pp. 828–841, 2013.

[6] L. Wang, J. Reiss, and A. Cavallaro, “Over-determined source separation and localization using distributed microphones,” *IEEE Trans. on Audio, Speech and Lang. Process.*, vol. 24, no. 9, pp. 1569–1584, 2016.

[7] T. Kim, H. Attias, S. Lee, and T. Lee, “Blind source separation exploiting higher-order frequency dependencies,” *IEEE Trans. on Audio, Speech and Lang. Process.*, vol. 15, no. 1, pp. 70–79, 2007.

[8] I. Lee, T. Kim, and T. Lee, “Fast fixed-point independent vector analysis algorithms for convolutive blind source separation,” *Signal Processing*, vol. 87, no. 8, pp. 1859–1871, 2007.

[9] T. Kim, “Real-time independent vector analysis for convolutive blind source separation,” *IEEE Trans. on Circuits and Systems*, vol. 57, pp. 1431–1438, 2010.

[10] J. Chen, J. Benesty, and Y. Huang, “Time delay estimation in room acoustic environments: an overview,” *EURASIP J. on App. Signal Process*, vol. 26503, no. 1, pp. 1–19, 2006.

[11] A. Brutti, A. Tsiami, A. Katsamanis, and P. Maragos, “A phase-based time-frequency masking for multi-channel speech enhancement in domestic environments,” *Interspeech*, pp. 2875–2879, 2016.

[12] T. Yoshioka, M. Delcroix, T. Nakatani, S. Araki, and M. Fujimoto, “Dominance based integration of spatial and spectral features for speech enhancement,” *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 21, no. 12, pp. 2516–2531, 2013.

[13] T. Yoshioka, T. Nakatani, M. Miyoshi, and H. G. Okuno, “Blind separation and dereverberation of speech mixtures by joint optimization,” *IEEE Trans. on Audio, Speech and Lang. Process.*, vol. 19, no. 1, pp. 69–84, 2011.

[14] M. Cobos and J. Lopez, “Maximum a posteriori binary mask estimation for underdetermined source separation using smoothed posteriors,” *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 20, no. 7, pp. 2059–2064, 2012.

[15] D. Wang, “Time-frequency masking for speech separation and its potential for hearing aid design,” *Trends in Amplification*, vol. 12, no. 4, pp. 332–353, 2008.

[16] S. Yan and C. Hou, “Broadband DOA estimation using optimal array pattern synthesis technique,” *IEEE Antennas Wirel. Propag. Lett.*, vol.5, pp. 88–90, March 2006.

[17] N. Roman, D. Wang, and G. Brown, “Speech segregation based on sound localization”, *J. Acoust. Soc. Amer.* vol. 114, no. 4, pp. 2236–2252, 2003.

[18] H. Sawada, S. Araki, and S. Makino, “Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment,” *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 19, no. 3, pp. 516–527, 2011.

[19] M. Mandel, R. Weiss, and D. Ellis, “Model-based expectation-maximization source separation and localization,” *IEEE Trans. on Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 382–394, 2010.

[20] S. Yan, Y. Ma, and C. Hou, “Optimal array pattern synthesis for broadband arrays,” *J. Acoust. Soc. Am.*, vol. 122, no. 5, pp. 2686–2696, Nov. 2007.

[21] Z. Huang, G. Zhan, D. Ying, and Y. Yan, “Robust multiple speech source localization using time delay histogram,” *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Shanghai, China, pp. 3191–3195, 2016.

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

[23] S. Rickard, R. Balan, and J. Rosca, “Real-time time-frequency based blind source separation,” *Proc. Int. Workshop Independent Component Anal. Blind Source Separation*, San Diego, CA, pp. 651–656, 2001.

[24] A. Joujine, S. Rickd, and O. Yılmaz, “Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures,” Proc. of *IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Istanbul, Turkey*, pp. 2985–2988, 2000.

[25] 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.

[26] Z. Huang, G. Zhan, D. Ying, and Y. Yan, “Robust Localization of Single Sound Source Based on Phase Difference Regression”, *Interspeech*, pp. 3293–3297, 2015.

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

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

[29] 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.

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