Window-Dominant Signal Subspace Methods for Multiple Short-Term Speech Source Localization

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

***Abstract*—Signal subspace has been widely exploited to localize multiple speech sources. However, most signal subspace methods cannot count the number of sources, and do not make use of speech sparsity in the frequency domain. This paper presents a grid search window-dominant signal subspace (GS-WDSS) method and a closed-form WDSS (CF-WDSS) method to localize short-term speech sources. Such methods are based upon the generalized sparsity assumption that each window containing some time-adjacent bins is dominated by one source, as opposed to the conventional assumption that each individual bin is dominated by one source. The generalized assumption enables the principal eigenvector of the spatial correlation matrix on each window to span the signal subspace of the window-dominant source. The direction-of-arrival (DOA) of the dominant source is estimated from the principal eigenvector. The DOAs and the number of sources are eventually summarized from the DOA histogram of all dominant sources. The conventional assumption is a special case of the generalized assumption. By using the generalized assumption, the performance in estimating DOAs of the window-dominant sources is significantly improved at the cost of acceptable masking effect. The superiority of the proposed methods is verified by simulated and real experiments.**

***Index Terms*—Speech source localization, signal subspace, speech sparsity, closed-form solution, window-dominant source.**

I. Introduction

Real-time speech source localization using microphone arrays is of great significance in numerous applications such as speech separation, speech enhancement, or speaker tracking [1]–[6]. The timely knowledge about speakers' locations is an essential prerequisite for these systems. Especially, in some scenarios where speech sources are present for a short time or their locations are time-varying, the localization must be conducted on a short-term segment such as a word-length utterance, instead of a long-term segment such as a sentence-length utterance. The challenging issue for short-term source localization is the robustness with respect to acoustic interferences such as environmental noises or reverberations. Generally speaking, the requirements to robust localization are twofold. One is the enough frames that are provided by the long-term source signal. The other is the robustness of localization method that resists interferences. As insufficient frames are available in a short-term segment, the robust localization method is required for the short-term source localization.

Signal subspace methods are robust for the short-term source localization. Numerous broadband subspace methods [9]–[24] that are derived from the narrowband subspace methods [7], [8] have been proposed in the past three decades. These methods have an inherent advantage to localize multiple broadband sources. The incoherent subspace methods (ISSM) are simple but effective to localize multiple speech sources with high spatial resolution. In these methods, the broadband speech signal is decomposed into multiple narrowband signals, and a narrowband subspace method is applied to each of these narrowband signals [9]–[11]. Contrary to ISSM, the coherent signal subspace methods (CSSM) coherently transform the spatial correlation matrix from each narrowband component to a reference narrowband, and then apply a narrowband subspace method on the reference narrowband [12]–[16]. Although CSSM was reported to outperform ISSM [12], CSSM depends on an initial focusing angle and an inaccurate estimate of this angle causes performance degradation. The covariance matrix based methods are closely related to the subspace methods, where the covariance matrix is expressed as the function of the steered direction [17]–[19]. When the array is steered toward the true direction of a source, the principal eigenvector spans the signal subspace of the steered source and the remaining eigenvectors span the subspace of noise and other sources.

Although the conventional subspace methods well address the problem of multiple sources, they still have three common disadvantages. First, most methods require the number of sources to be estimated in advance, and are sensitive to the estimate of the source number. Second, their computational efficiency is undesirable due to the widely-used grid search. Estimation of signal parameters via rotational invariance techniques (ESPRIT) is a special subspace method that does not require grid search. But ESPRIT was originally designed for the uniform linear array [8], [20], [21] and the complicated transformation from the arbitrary array geometry into the linear geometry results in reduction of estimation accuracy [22]–[24]. Even though the grid search can be accelerated by optimizing the search strategy, the acceleration is achieved at the cost of performance degradation [25]–[28]. Lastly, the speech signal is commonly taken as a general broadband signal, but such an approach ignores a special property of speech signals.

Speech is a special broadband signal which is sparsely distributed in the frequency domain [29], [30]. It is conventionally assumed that multiple speech sources have different distribution in the frequency domain, and that at most one source is dominant in power for an individual bin while the contributions from the remaining speech sources are negligible. This sparsity assumption is referred to as W-disjoint orthogonality in [29]. Such an assumption can simplify the

The authors are with the key laboratory of speech acoustics and content understanding, Institute of Acoustics, Chinese Academy of Sciences, Beijing, Haidian District, 100190 China. e-mail: yingdongwen@hccl.ioa.ac.cn and yyan@hccl.ioa.ac.cn.

multiple source localization on the full frequency band to the single source localization at individual bins. On linear arrays, numerous closed-form methods have been presented based on a simplified model where direction-of-arrival (DOA) estimation is equivalent to the time delay estimation [31]–[39]. On planar arrays, the simplified model enables some closed-form methods [40]–[42], whereas some methods still use the grid search to find DOAs due to their complicated cost function [43]–[47]. The closed-form methods usually exhibit high computational efficiency as the grid search is avoided.

Robustness as well as computational efficiency are important factors for short-term speech source localization. For long-term speech sources, the poor robustness of localization method may not substantially degrade the performance because the interferences can be effectively alleviated by sufficient frames. Most sparsity-based methods were reported to work on long-term speech segments with duration of at least one second [29], [32]–[35], [37]–[46]. If those methods are applied on the short-term speech sources with duration much less than one second, the acoustic interferences become the major hindrance for robust localization.

By making use of the speech sparsity and the signal subspace, this paper proposes two real-time methods to estimate DOAs of multiple sources on short-term speech segments. The contributions of this paper are twofold. First, this paper makes the generalized sparsity assumption that each window containing some time-adjacent bins is dominated by at most one speech source. The conventional sparsity assumption [29] is a special case of the generalized assumption. Second, this paper relates the signal subspace to the generalized assumption. A grid search subspace method and a closed-form subspace method are respectively presented based on this relation.

The remainder of this paper is organized as follows: Section II formulates the signal model and reviews the related works. Section III investigates the positive and negative effects of the generalized assumption, and presents a grid search method. Section IV presents a closed-form method based on the generalized assumption. Section V gives the details of implementation. Section VI gives various experiments and discussions, and Section VII concludes the study.

II. Preliminaries

A. Signal model

The signal model considers the far-field scenario that D speech sources impinge on a K-element planar array. Assuming the free-field model, the signal received by the kth microphone is expressed as

$$x_k(t) = \sum_{d=1}^{D} s_d(t - \psi_{k,d}) + n_k(t), \tag{1}$$

where $s_d(t)$ denotes the signal emitted from the dth source, $n_k(t)$ denotes the additive noise at the kth microphone which is uncorrected with the source signal, and $\psi_{k,d}$ denotes the propagation time from source d to microphone k. By applying the short-time Fourier transform (STFT) to (1), we obtain

$$X_k(\omega_i) = \sum_{d=1}^{D} S_d(\omega_i) e^{-j\omega_i \psi_{k,d}} + N_k(\omega_i), \tag{2}$$

where ω_i is the angular frequency at the ith frequency, j is an imaginary unit, $S_d(\omega_i)$ denotes the Fourier coefficient of the source signal, and $N_k(\omega_i)$ denotes the Fourier coefficient of the noise signal. Using vector notation, (2) is re-written as

$$\mathbf{x}(\omega_i) = \sum_{d=1}^{D} \mathbf{a}_i S_d(\omega_i) + \mathbf{n}(\omega_i), \tag{3}$$

where

$$\mathbf{x}(\omega_i) = \Big[X_1(\omega_i), \cdots, X_K(\omega_i)\Big]^T,$$
$$\mathbf{n}(\omega_i) = \Big[N_1(\omega_i), \cdots, N_K(\omega_i)\Big]^T,$$
$$\mathbf{a}_i = \Big[e^{-j\omega_i\psi_{1,d}}, e^{-j\omega_i\psi_{2,d}}, \cdots, e^{-j\omega_i\psi_{K,d}}\Big]^T,$$

where $\mathbf{a}_i$ is the steering vector for the dth source. If the first microphone is used as the reference, the steering vector is represented as

$$\mathbf{a}_i = e^{-j\omega_i\psi_{1,d}} \times \Big[1, e^{-j\omega_i(\psi_{2,d}-\psi_{1,d})}, \cdots, e^{-j\omega_i(\psi_{K,d}-\psi_{1,d})}\Big]^T. \tag{4}$$

Any microphone can be taken as the reference. The steering vector is determined by the propagation time delay between two microphones. The K-element array consists of M microphone pairs. Given the dth source, the time delay between the mth microphone pair is represented as the function of its DOA, as illustrated by Fig. 1(b). It is expressed as

$$\tau_m^{(d)} = \mathbf{g}_m^T \boldsymbol{\gamma}^{(d)} r_m / c, \tag{5}$$

where c denotes the sound velocity, r_m denotes the distance between the two microphones, and the unit vector $\boldsymbol{\gamma}^{(d)}$ denotes the DOA of the dth source, which is given by

$$\boldsymbol{\gamma}^{(d)} = \Big[\cos\alpha^{(d)}\cos\beta^{(d)},\ \sin\alpha^{(d)}\cos\beta^{(d)},\ \sin\beta^{(d)}\Big]^T, \tag{6}$$

where $\alpha^{(d)}$ denotes the azimuth angle, $\beta^{(d)}$ denotes the elevation angle, and $(.)^T$ denotes the transpose. The direction vector between the locations of the mth pair of microphones is expressed by a unit vector,

$$\mathbf{g}_m = \Big[g_{m,1}, \quad g_{m,2}, \quad 0\Big]^T, \tag{7}$$

where the third dimension is set to zero, signifying that all microphones are arranged in a plane. Eventually, the steering vector is represented as the function of the direction, which is denoted as $\mathbf{a}_i(\boldsymbol{\gamma}^{(d)})$ in the followings. The purpose of sound source localization is to estimate the DOAs of D sources from the received signal $\mathbf{x}(\omega_i)$.

B. Broadband MUSIC

The broadband multiple signal classification (MUSIC) is a classical signal subspace method based on the multi-source model in (3). It utilizes the orthogonality between the steering vector and the noise subspace. $N_k(\omega_i)$ is assumed to be additive white Gaussian noise. The noise subspace is obtained by decomposition of the time-averaged spatial correlation matrix at each frequency, which is given by

$$\mathbf{R}_i = E\Big[\mathbf{x}(\omega_i)\mathbf{x}^H(\omega_i)\Big] = \mathbf{A}_i \mathbf{R}_i^{(s)} \mathbf{A}_i^H + \sigma_i^2 \mathbf{I}, \tag{8}$$

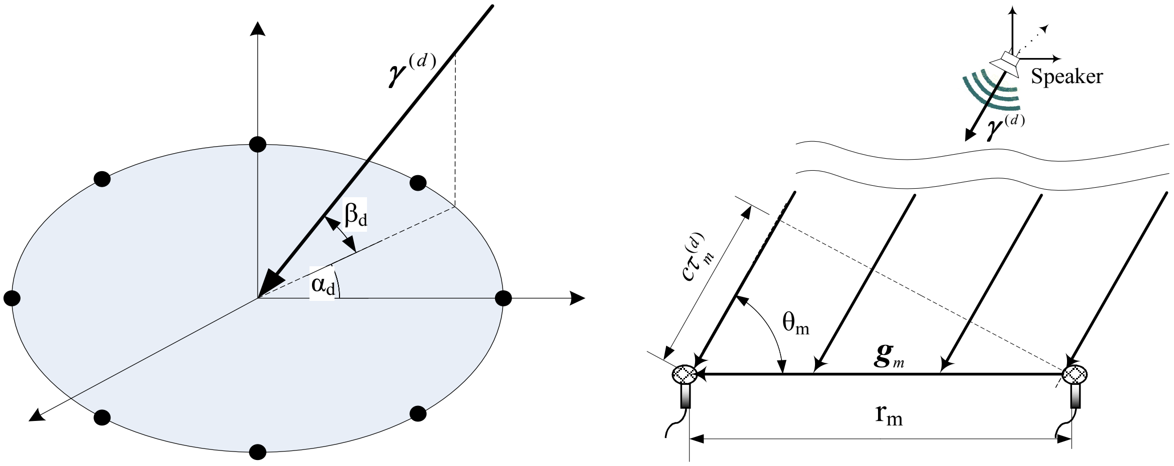

Fig. 1. Geometrical relationship: (a) Impinging direction for the dth source; (b) Time delay for the mth pair of microphones.

where $(.)^H$ denotes the conjugate transpose, $E(.)$ denotes the expectation over time, and

$$\begin{aligned} \mathbf{R}_i^{(s)} &= E\big[\mathbf{s}_i\mathbf{s}_i^H\big], \\ \mathbf{A}_i &= \Big[\mathbf{a}_i(\boldsymbol{\gamma}^{(1)}), \mathbf{a}_i(\boldsymbol{\gamma}^{(2)}), \cdots, \mathbf{a}_i(\boldsymbol{\gamma}^{(D)})\Big], \\ \mathbf{s}_i &= \Big[S_1(\omega_i), S_2(\omega_i), \cdots, S_D(\omega_i)\Big]. \end{aligned} \tag{9}$$

The $D \times D$ matrix $\mathbf{R}_i^{(s)}$ has full rank if D sources are mutually uncorrelated, and then the noise subspace is formed from the eigenvalue decomposition of the correlation matrix. The eigenmatrix for the noise subspace is expressed as

$$\mathbf{V}_i = \Big[\mathbf{u}_{i,D+1}, \cdots, \mathbf{u}_{i,K}\Big], \tag{10}$$

where the K eigenvectors are arranged in the descending order with respect to their eigenvalues.

The cost function of the broadband MUSIC tests the orthogonality between the steering vector and the noise subspace on all available frequencies, which is given by

$$f_{MUSIC}(\boldsymbol{\gamma}) = \sum_i \mathbf{a}_i^H(\boldsymbol{\gamma})\mathbf{V}_i\mathbf{V}_i^H\mathbf{a}_i(\boldsymbol{\gamma}). \tag{11}$$

The DOAs of the multiple sources are estimated by minimizing the cost function. Unfortunately, (11) is a trigonometric function, which does not have a closed-form solution to DOA. A grid search has to be used to find the optimal estimates of DOAs. Most signal subspace methods rely on an accurate estimate of the source number to discriminate the noise subspace from the signal subspace. Properly speaking, the source number at each frequency should be accurately estimated in advance since the speech signal is sparsely distributed in the frequency domain. For long-term speech segments, the source number at each frequency is generally equivalent to the number at the full band since the sufficient frames can guarantee that all speech sources are present at every frequency. However, this equivalence does not hold true for short-term segments because the speech sources are usually absent in some frequencies. The source number at each frequency ranges from zero to the maximum (e.g. the source number at the full band). It is quite difficult to accurately estimate the number of sources on individual frequencies for a short time. The ambiguity in the source number causes difficulty when applying the subspace methods on the short-term speech sources. Fortunately, this problem can be avoided by considering the speech sparsity.

C. Sparse source localization

Speech is usually taken as a sparse source signal in numerous localization methods [29]–[47]. Based on speech sparsity, the multi-source model (3) is simplified to a single-source model,

$$\mathbf{x}(\omega_i) \approx \mathbf{a}_i(\boldsymbol{\gamma}^{(d')})S_{d'}(\omega_i), \tag{12}$$

where the index for the dominant source is given by

$$d' = \arg\max_{d\in[1:D]} \Big|S_d(\omega_i)\Big|. \tag{13}$$

Most sparsity-based methods regard the background noise as a nondirectional source, and do not take into account noise-dominating bins. Therefore, the noise item is disregarded in the signal model.

The time delays that are derived from the phase difference of Fourier coefficients are widely employed to estimate the DOA of the dominant source [29]–[40], [42]. For a given bin, the time delay for the dominant source is given by

$$\tau_{m,i}^{(d')} \approx \frac{\angle X_{k_2}(\omega_i) - \angle X_{k_1}(\omega_i)}{\omega_i} + pT_i, \tag{14}$$

where $\angle(.)$ denotes the phase operation, p denotes the period number, and T_i denotes the period at the ith frequency, given by

$$T_i = 2\pi/\omega_i.$$

In all potential delays, there exists the minimal delay, $\eta_{m,i}^{(d')} \in [-T_i/2, T_i/2]$. The time delay is expressed as the sum of the minimal delay and $p_{m,i}$ periods, given by

$$\tau_{m,i}^{(d')} \approx \eta_{m,i}^{(d')} + p_{m,i}T_i. \tag{15}$$

The minimal delay can be easily determined from the phase difference. But the period number results in ambiguity in the time delay. The ambiguity is often resolved by the geometric constraint, as illustrated in Fig. 1(b). The constraint is given by

$$-r_m \le c\tau^{(d')}_{m,i} \le r_m. \tag{16}$$

The sufficiently small distance enables the minimal delay to be the unique candidate [39]–[41], [47], namely $\tau^{(d')}_{m,i} = \eta^{(d')}_{m,i}$. However, the small distance will favour the coherence in low frequencies [31], which degrades the performance of localization. Ambiguity about the period number is encountered when the time delay is estimated from widely spaced microphones. There may exist several candidates for a period number, which is given by the set,

$$P_{m,i} = \left\{ p \,\middle|\, \left\lceil \frac{-r_m - c\eta^{(d')}_{m,i}}{cT_i} \right\rceil \le p \le \left\lfloor \frac{r_m - c\eta^{(d')}_{m,i}}{cT_i} \right\rfloor \right\}, \tag{17}$$

where $\lceil . \rceil$ and $\lfloor . \rfloor$ respectively denote the ceil and floor integer operations. On linear arrays, some methods traverse all the candidates to find the optimal one [33], [37]. But the traversing method is very computationally expensive on the widely-spaced planar array. There are in total $\prod_{m=1}^{M} \mathcal{C}(p_{m,i})$ potential combinations for the period numbers, $[p_{1,i}, \cdots, p_{M,i}]$, where $\mathcal{C}(.)$ denotes the operation of taking cardinality of the set of all possible values. For long-term speech sources, the time-delay histogram is a simple but effective method to resolve the ambiguity [42]. But a short-term speech segment provides insufficient time delays to construct a reliable histogram. Estimation of the period number is a critical point in the time delay estimation.

For the mth microphone pair, the time delay can determine an included angle between the microphone pair direction and the DOA of the dominant source. As shown in Fig. 1(b), the geometric relationship is expressed as

$$\cos\theta_{m,i} = c\tau^{(d')}_{m,i}/r_m. \tag{18}$$

But the included angle has an ambiguity to DOA in a three-dimensional space. At a specific bin, an explicit DOA is determined by at least two time delays that are estimated from the unaligned microphones. For a planar array, the DOA of the dominant source is represented as the function of all the delays. Finally, the DOAs and the number of the speech sources are obtained by the histogram or clustering analysis from the DOAs of all dominant sources. The sparse source methods will have the advantage in computational efficiency if the closed-form solution can be derived from the simplified model. However, the disadvantage is the masking effect at each bin, where the dominant source masks the signals of the remaining sources. If a weak speech source is masked by a strong source at most bins, the weak source will be missdetected or inaccurately localized. The relationship between the dominant and masked sources is investigated thoroughly in [29].

III. Speech sparsity and grid search method

The dominant source is conventionally considered at individual bins, whereas this paper considers the dominant source in the window that contains L time-adjacent bins. It is well known that the spectral amplitude of a speech source is temporally correlated, and therefore, the dominant source at a given bin is usually identical to the dominant sources at its time-adjacent bins. Based on such fact, the sparsity assumption is generalized from each individual bin to the L-bin window.

We collect a window of Fourier coefficients at frequency bin i that originates from time ℓ up to $\ell + L - 1$, namely $\left[\mathbf{x}_\ell(\omega_i), \mathbf{x}_{\ell+1}(\omega_i), \cdots, \mathbf{x}_{\ell+L-1}(\omega_i)\right]$. The dominant source associates with the maximal intensity, the index of which is expressed as

$$d' = \arg\max_{d\in[1:D]} \sum_{t=\ell}^{\ell+L-1} \left| S_{d,t}(\omega_i) \right|^2. \tag{19}$$

With the sparsity assumption, an approximation is given by

$$\sum_{t=\ell}^{\ell+L-1} \left\| \mathbf{x}_t(\omega_i) \right\|^2 \approx \sum_{t=\ell}^{\ell+L-1} \left| S_{d',t}(\omega_i) \right|^2, \tag{20}$$

where the index for the window-dominant source is given by (19). Accordingly, the spatial correlation matrix on the (i, ℓ)th window is approximated as

$$\begin{aligned} \mathbf{R}_i &= \frac{1}{L} \sum_{t=\ell}^{\ell+L-1} \mathbf{x}_t(\omega_i)\mathbf{x}_t^H(\omega_i) \\ &\approx \mathbf{a}_i(\boldsymbol{\gamma}^{(d')})\mathbf{a}_i^H(\boldsymbol{\gamma}^{(d')}) \cdot \sum_{t=\ell}^{\ell+L-1} \frac{\left| S_{d',t}(\omega_i) \right|^2}{L}. \end{aligned} \tag{21}$$

The approximation means $\mathbf{R}_i$ to be a rank-1 matrix, which enables the steering vector of the dominant source to be approximated as the principal eigenvector of the signal subspace. It is expressed as

$$\begin{aligned} &\mathbf{u}_{i,1} \approx e^{-j\omega_i z_i} \mathbf{a}_i(\boldsymbol{\gamma}^{(d')}) / \left\| \mathbf{a}_i(\boldsymbol{\gamma}^{(d')}) \right\|, \\ &\angle u_{i,1,k} = -\omega_i(\psi_{k,d'} + z_i), \end{aligned} \tag{22}$$

where $u_{i,1,k}$ is the kth element of principal eigenvector $\mathbf{u}_{i,1}$, and z_i is the arbitrary real constant that is introduced by the complex eigenvalue decomposition.

The approximation in (20) is equivalent to that in (22). They are evaluated by the similarity between the principal eigenvector and the steering vector, which is defined as

$$\rho_{d',i} = \frac{\left| \mathbf{u}_{i,1}^H \mathbf{a}_i(\boldsymbol{\gamma}^{(d')}) \right|}{\left\| \mathbf{a}_i(\boldsymbol{\gamma}^{(d')}) \right\|}. \tag{23}$$

The similarity degree ranges from a completely uncorrelated 0 to a completely similar 1. If the contribution from the masked sources is very small relative to the contribution from the dominant source, the similarity degree will be close to 1. The similarity degree reflects to what extent the generalized assumption holds true. Several simulated experiments were conducted to investigate the generalized assumption.

The first experiment investigated the influence of the window size on the similarity degree. An eight-element circular array was placed with horizontal orientation in the center of a simulated room with dimensions of $5 \times 8 \times 2.7$ m. Most short-term speech segments have no more than three sources in a common sense because listeners are unable to correctly percept the target source from more than three competing

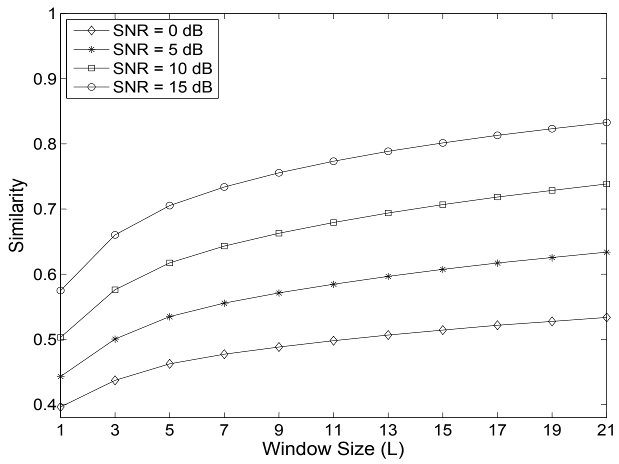

Fig. 2. Similarity degree under various window sizes for the anechoic condition.

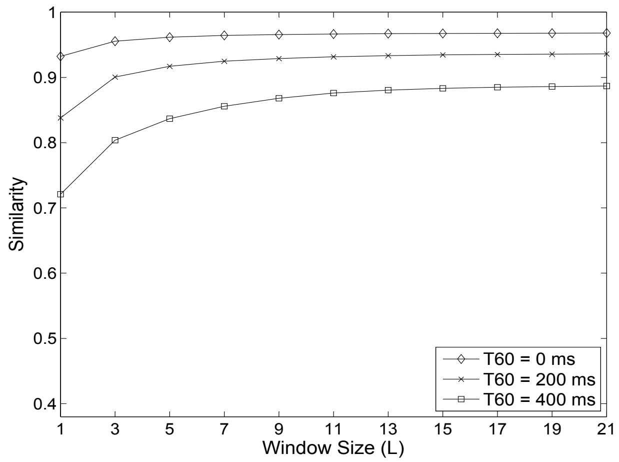

Fig. 3. Similarity degree under various window sizes for the noise free condition.

speakers in a phrase-length utterance [48]. Therefore, all experiments were set to at most three sources in this paper. Three long utterances with a duration of 200 seconds were respectively collected from TIMIT database [49], and taken as the speech sources. They were respectively set at the azimuth angles of $121°$, $177°$, and $236°$, and at a distance of 1.2 m from the array center. The environments were simulated by the image source method [50]–[52] under various reverberation times. The real noise was recorded in a room by the circular array. It was added to the simulated data with various SNRs. All experiments employed 256-point FFT and 32-ms frames with a half-frame overlap. The sampling rate of all signals is 8 kHz. Fig. 2 and Fig. 3 plot the relationship between the window size and the similarity under an anechoic condition and a noise free condition, respectively. This experiment shows the large window size is helpful to improve the similarity degree. The performance of DOA estimation for the window-dominant source is highly correlated with the similarity degree. Compared with the conventional assumption ($L = 1$), the generalized assumption ($L > 1$) significantly improves the robustness with respect to acoustic interferences.

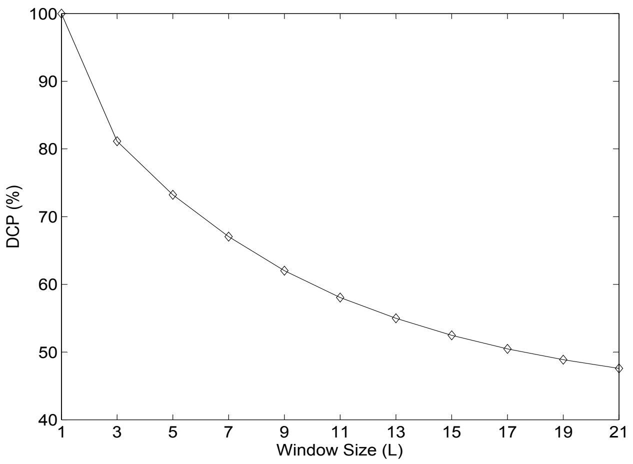

Fig. 4. DCP under various window sizes.

The generalized assumption leads to the more serious masking effect than the conventional assumption [29]. The window-dominant source masks the remaining sources even if the remaining sources dominate some bins inside the window. The masking effect is assessed by using the consistency between the window-dominant sources and the bin-dominant sources. The index for the (i, ℓ)-th window is denoted as $d'_{i,\ell}$, and the indices of the bin-dominant sources inside the window are denoted as $\left[d'^{(1)}_{i,\ell}, d'^{(2)}_{i,\ell}, \cdots, d'^{(L)}_{i,\ell}\right]$. A dominance consistency probability (DCP) describes to what extent the bin-dominant sources are consistent with the window-dominant sources. DCP is defined as

$$DCP = \frac{\sum_{\ell}\sum_{i}\sum_{t=1}^{L}\delta(d'_{i,\ell} - d'^{(t)}_{i,\ell})}{\sum_{\ell}\sum_{i}\sum_{t=1}^{L}\delta(0)} \times 100\%, \qquad (24)$$

where δ is the Dirac function. A large value of DCP means the high consistency that validates the generalized assumption. The inconsistency leads to the masking effect, and $100\% - DCP$ denotes the percentage of the masked bins.

The second experiment investigated the relationship between DCP and the window size. The experimental setup was the same as the setup used in the first experiment. But the noise and reverberation were not considered here because they did not change the conclusion. For a given window size, DCP is averaged over all times and all available frequencies, as described by (24). Fig. 4 plots DCPs under various sizes. The result is twofold. First, most bin-dominant sources are consistent with the window-dominant source for small-size windows since DCP is generally much larger than chance probability 33.3%. This result verifies the correctness of the generalized assumption on small-size windows. Second, the large window size enhances the masking effect, and the conventional assumption ($L = 1$) has the minimal masking effect. It is found out from Figs. 2–4 that the 5-frame window makes a good tradeoff between the masking effect and the robustness for the dominant source DOA estimation. Therefore, the window size is set to 5 frames in all following experiments.

Using the five-frame window, we investigated the relationship between the similarity degree and the dominance degree

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2016.2625458, IEEE/ACM Transactions on Audio, Speech, and Language Processing

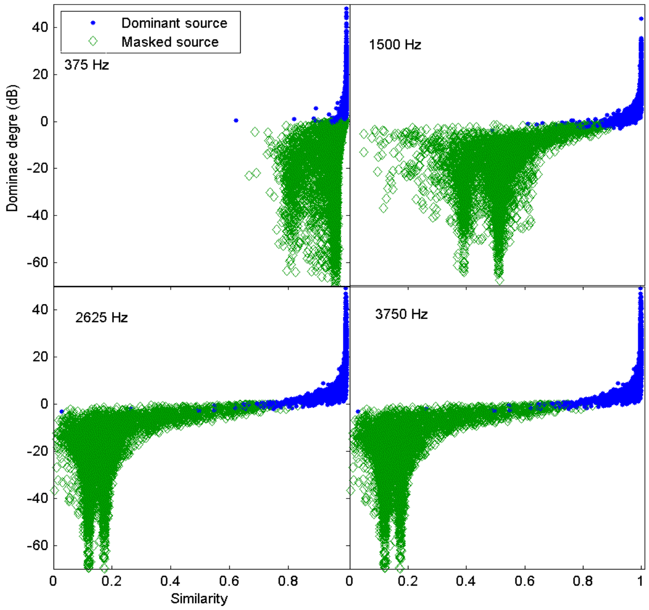

Fig. 5. Scatter plots between similarity and dominance degrees.

on the aforementioned simulation. For the d'th source at the (i,ℓ)-th window, the dominance degree is defined as

$$\kappa_{d',i} = 10\log_{10}\left[\frac{\sum_{t=\ell}^{\ell+L-1}\left|S_{d',t}(\omega_i)\right|^2}{\sum_{\substack{d=1\\d\neq d'}}^{D}\sum_{t=\ell}^{\ell+L-1}\left|S_{d,t}(\omega_i)\right|^2}\right]. \qquad (25)$$

Fig. 5 shows scatter plots of points (ρ,κ), at different frequencies, wherein the experimental setup is the same as that in the second experiment. The blue points denote the dominant sources with the large ratio, and the green points denote the masked sources with the small ratio. This figure shows that the dominant sources generally associate with the high similarity, which validates the approximation in (20) and (22). At low frequencies, the similarity discrimination over the dominance degree is insignificant because of the spatial coherence.

By means of testing the similarity degree, a grid search window-dominant source signal subspace method (GS-WDSS) is proposed to estimate the DOA of the window-dominant source. The DOA $\hat{\gamma}_{i,\ell}^{(d')}$ is determined by maximizing the similarity in Algorithm 1, where the subscript $(.)^{(d')}$ and $(.)_\ell$ are omitted for simplicity in the following. The computational efficiency of this algorithm remains a problem due to the grid search.

IV. Closed-Form Method

This section presents a closed-form window-dominant source signal subspace method (CF-WDSS) with high computational efficiency. The DOA estimator for the window-dominant source is derived from the single-source method in [53]. But the purpose of CF-WDSS is multiple source localization. CF-WDSS addresses not only acoustic interferences, but also ambiguity about the period number in time delay estimation. Fig. 6 plots the block diagram of the CF/GS methods, where $h-1$ low frequencies are disregarded because of spatial coherence.

Algorithm 1 DOA estimation of the dominant source using GS-WDSS

```
Estimate the correlation matrix at a window using (8);
Perform eigenvalue decomposition and extract the principal eigenvector;
for each $\alpha \in [0°, 360°)$ do
    for each $\beta \in [0°, 90°]$ do
        Calculate the test DOA using (6);
        Calculate the test time delays using (5);
        Calculate the test steering vector using (4);
        Calculate the similarity degree using (23);
    end for
end for
Take the maximal-similarity candidate as the direction.
```

A. DOA estimation of dominant sources

For CF-WDSS, the DOA of the window-dominant source is estimated from the time delays. Even though the propagation times can not be obtained due to the uncertainty of constant z_i, the time delays can be estimated from the phase differences of the principal eigenvector in CF-WDSS, as opposed to the time delays that are estimated from Fourier coefficients in conventional methods [29], [31]–[42]. The time delay is given by

$$\tau_{m,i} = \frac{\angle u_{i,1,k_2} - \angle u_{i,1,k_1}}{\omega_i} + pT_i. \qquad (26)$$

With the same principle in (15), the delay for CF-WDSS is represented by the summation of minimal delay $\eta_{m,i}$ and a certain number of periods $p_{m,i}T_i$. The time delay can not be obtained until the ambiguity in the period number is resolved. Nevertheless, the closed-form solution to DOA is given before the solution to the period number because the latter depends on the former. The latter will be addressed in the next subsection.

The cosine of included angle $\theta_{m,i}$ is estimated from the time delay, which is expressed as

$$\cos\hat{\theta}_{m,i} = c(\eta_{m,i} + p_{m,i}T_i)/r_m, \qquad (27)$$

By utilizing the geometric relationship, the estimate of this cosine is expressed as

$$\cos\hat{\theta}'_{m,i} = \mathbf{g}_m^T\boldsymbol{\gamma}_i. \qquad (28)$$

In desirable conditions, the difference between the two cosines should be zero. But there exists an error between them in the presence of acoustic interferences. At a given window, the error function is defined as the weighted square error of cosines of all included angles across all microphone pairs. It is given by

$$\begin{aligned} f_i(\boldsymbol{\gamma}_i) &= \sum_{m=1}^{M} w_{m,i}\left[\cos\hat{\theta}_{m,i} - \cos\hat{\theta}'_{m,i}\right]^2 \\ &= \sum_{m=1}^{M} w_{m,i}\left[\mathbf{g}_m^T\boldsymbol{\gamma}_i - c(\eta_{m,i} + p_{m,i}T_i)/r_m\right]^2, \end{aligned} \qquad (29)$$

where $w_{m,i}$ is the coefficient that weights the reliability of the mth microphone pair, and $\mathbf{g}_m$ is given by (7). By means

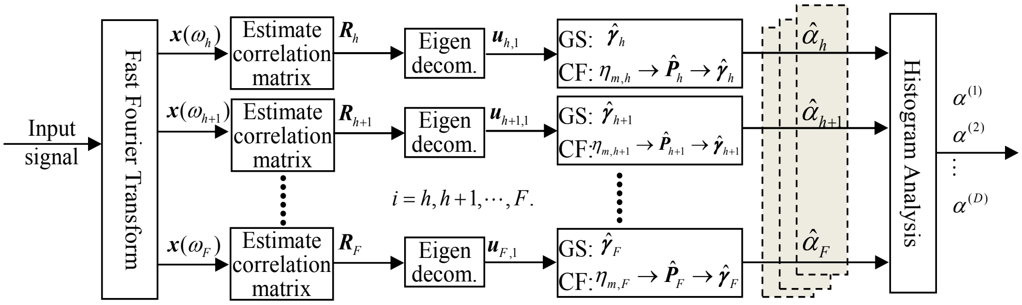
Input
signal
Fast Fourier Transform
$\boldsymbol{x}(\omega_h)$
$\boldsymbol{x}(\omega_{h+1})$
$\boldsymbol{x}(\omega_F)$
Estimate correlation matrix
Estimate correlation matrix
Estimate correlation matrix
$\boldsymbol{R}_h$
$\boldsymbol{R}_{h+1}$
$\boldsymbol{R}_F$
Eigen decom.
Eigen decom.
Eigen decom.
$\boldsymbol{u}_{h,1}$
$\boldsymbol{u}_{h+1,1}$
$\boldsymbol{u}_{F,1}$
$i = h, h+1, \cdots, F.$
GS: $\hat{\boldsymbol{\gamma}}_h$
CF: $\eta_{m,h} \rightarrow \hat{\boldsymbol{P}}_h \rightarrow \hat{\boldsymbol{\gamma}}_h$
GS: $\hat{\boldsymbol{\gamma}}_{h+1}$
CF: $\eta_{m,h+1} \rightarrow \hat{\boldsymbol{P}}_{h+1} \rightarrow \hat{\boldsymbol{\gamma}}_{h+1}$
GS: $\hat{\boldsymbol{\gamma}}_F$
CF: $\eta_{m,F} \rightarrow \hat{\boldsymbol{P}}_F \rightarrow \hat{\boldsymbol{\gamma}}_F$
$\hat{\alpha}_h$
$\hat{\alpha}_{h+1}$
$\hat{\alpha}_F$
Histogram Analysis
$\alpha^{(1)}$
$\alpha^{(2)}$
$\vdots$
$\alpha^{(D)}$

Fig. 6. Block diagram of the proposed methods.

of minimizing the error, the closed-form solution to DOA is given by

$$\begin{gathered}\hat{\gamma}_i = \min_{\boldsymbol{\gamma}} f_i(\boldsymbol{\gamma}) \\ \text{subjected to: } \boldsymbol{\gamma}^T\boldsymbol{\gamma} = 1.\end{gathered} \tag{30}$$

By making use of the Karush-Kuhn-Tucker necessary conditions, an optimal estimate is given by (31), where

$$\begin{aligned}\mathbf{g}'_m &= \begin{bmatrix} g_{m,1}, & g_{m,2}\end{bmatrix}^T, \\ \hat{\boldsymbol{\gamma}}_i &= \begin{bmatrix}\hat{\gamma}_{1,i}, & \hat{\gamma}_{2,i}, & \hat{\gamma}_{3,i}\end{bmatrix}^T.\end{aligned}$$

The weight coefficient is determined by the error between two angles which are obtained by applying the inverse cosine function on (27) and (28). The error is given by

$$\delta_{m,i} = \hat{\theta}_{m,i} - \hat{\theta}'_{m,i}. \tag{33}$$

Assuming that the error conforms a zero-mean Gaussian distribution with variance $\sigma_i^2 = \sum_{m=1}^{M} \delta_{m,i}^2/M$, the weight is represented by the normalized likelihood. It is expressed as

$$w_{m,i} = \exp(-\delta_{m,i}^2/\sigma_i^2) / \sum_{m=1}^{M} \exp(-\delta_{m,i}^2/\sigma_i^2). \tag{34}$$

The unreliable estimates of the time delays generally lead to large error, and thereby associate with small weights, mitigating negative effects.

B. Estimation of period number

Estimation of the period number is a key point to resolve the optimal estimate of the dominant source's DOA. The criterion of resolving the period number, $\mathbf{P}_i = \left[p_{1,i}, p_{2,i}, \cdots, p_{M,i}\right]$, is to minimize the cost function in (29). The closed-form solution to the period number is derived by differentiating the cost function with respect to $\mathbf{P}_i$ and solving for its zero.

Substituting the solution (31) in place of the unit direction in (29), we obtain (32) as the cost function for resolving the period number, where

$$\mathbf{Z} = \Big[\sum_{m=1}^{M} \mathbf{g}'_m \mathbf{g}'^T_m\Big]^{-1}, \tag{35}$$

and all weight coefficients are set as 1 for the purpose of high computational efficiency. The optimal estimates of the period numbers are given by minimizing the cost function, which is expressed as

$$\left[\hat{p}_{1,i}, \hat{p}_{2,i}, \cdots, \hat{p}_{M,i}\right] = \min_{\left[p_1, p_2, \cdots, p_M\right]} f'_i(p_1, p_2, \cdots, p_M). \tag{36}$$

The cost function is convex with respect to $\mathbf{P}_i$, and therefore has a global minimum. We take the first-order derivative of (32) with respect to $p_{h,i}$, and set it to zeros, which is expressed as

$$\begin{aligned}\frac{\partial f'_i}{\partial p_{h,i}} =& T_i \sum_{m=1, m\neq h}^{M} p_{m,i}\Big[\mathbf{C}_h\mathbf{g}'_m - \mathbf{g}'^T_m\mathbf{Z}\mathbf{g}'_h + v_h\mathbf{g}'^T_h\mathbf{Z}\mathbf{g}'_m\Big]\Big/ r_m \\ &+ p_{h,i}T_i\Big[\mathbf{C}_h\mathbf{g}'_h + v_h^2\Big]/r_h + \mathbf{C}_h\mathbf{Q}_i + v_h\mathbf{g}'^T_h\mathbf{Z}\mathbf{Q}_i \\ &- \sum_{m=1, m\neq h}^{M} \mathbf{g}'^T_m\mathbf{Z}\mathbf{g}'_h\eta_{m,i}/r_m - v_h\eta_{h,i}/r_h = 0,\end{aligned} \tag{37}$$

where

$$\mathbf{C}_h = \sum_{m=1, m\neq h}^{M} \mathbf{g}'^T_m\mathbf{Z}\mathbf{g}'_h\mathbf{g}'^T_m\mathbf{Z}, \tag{38}$$

$$\mathbf{Q}_i = \sum_{m=1}^{M} \eta_{m,i}\mathbf{g}'_m/r_m, \tag{39}$$

$$v_h = \mathbf{g}'^T_h\mathbf{Z}\mathbf{g}'_h - 1. \tag{40}$$

The coefficients of $p_{m,i}$ in (37) are combined as

$$\Gamma_{h,m} = \begin{cases}\Big[\mathbf{C}_h\mathbf{g}'_m - \mathbf{g}'^T_m\mathbf{Z}\mathbf{g}'_h + v_h\mathbf{g}'^T_h\mathbf{Z}\mathbf{g}'_m\Big]\Big/ r_m & m \neq h \\ \Big[\mathbf{C}_m\mathbf{g}'_m + v_m^2\Big]\Big/ r_m & m = h.\end{cases} \tag{41}$$

The constant items in (37) are combined as

$$\begin{aligned}\Upsilon_{h,i} =& \sum_{m=1, m\neq h}^{M} \mathbf{g}'^T_m\mathbf{Z}\mathbf{g}'_h\eta_{m,i}/r_m + v_h\eta_{h,i}/r_h \\ &- \mathbf{C}_h\mathbf{Q}_i - v_h\mathbf{g}'^T_h\mathbf{Z}\mathbf{Q}_i.\end{aligned} \tag{42}$$

The equation (37) can thus be rewritten as

$$\sum_{m=1}^{M} \Gamma_{h,m}\hat{p}_{m,i} = \Upsilon_{h,i}/T_i. \tag{43}$$

$$\begin{gathered}\left[\begin{array}{c}\hat{\gamma}_{1,i}\\ \hat{\gamma}_{2,i}\end{array}\right]=\left[\sum_{m=1}^{M}w_{m,i}\mathbf{g}'_m\mathbf{g}'^T_m\right]^{-1}\sum_{m=1}^{M}cw_{m,i}(\eta_{m,i}+p_{m,i}T_i)\mathbf{g}'_m/r_m,\\ \hat{\gamma}_{3,i}=\sqrt{1-\hat{\gamma}^2_{1,i}-\hat{\gamma}^2_{2,i}}.\end{gathered}\tag{31}$$

$$f'_i(p_1,p_2,\cdots,p_M)=\sum_{m=1}^{M}\left[\mathbf{g}'^T_m\mathbf{Z}\sum_{q=1}^{M}\Big[c(p_qT_i+\eta_{q,i})\mathbf{g}'_q/r_q\Big]-c(p_mT_i+\eta_{m,i})/r_m\right]^2.\tag{32}$$

With the same principle, a group of equations for h ranging from 1 to M are obtained. These equations are summarized by a matrix equation,

$$T_i\mathbf{\Gamma}\hat{\mathbf{P}}_i=\mathbf{\Upsilon}_i,\tag{44}$$

where

$$\mathbf{\Gamma}=\left[\begin{array}{cccc}\Gamma_{11}&\Gamma_{12}&\cdots&\Gamma_{1M}\\ \Gamma_{21}&\Gamma_{22}&\cdots&\Gamma_{2M}\\ &\cdots&&\cdots\\ \Gamma_{M1}&\Gamma_{M2}&\cdots&\Gamma_{MM}\end{array}\right],\tag{45}$$

$$\mathbf{\Upsilon}_i=\left[\Upsilon_{1,i},\Upsilon_{2,i},\cdots,\Upsilon_{M,i}\right]^T.\tag{46}$$

Eventually, the estimate of the period numbers is given by

$$\hat{\mathbf{P}}_i=\mathbf{\Gamma}^{-1}\mathbf{\Upsilon}_i/T_i,\tag{47}$$

where $\mathbf{\Gamma}$ is the function of the array topology, irrelevant to frequency or time delays. For a given array, constant matrix $\mathbf{\Gamma}^{-1}$ is shared for all estimations. $\mathbf{\Upsilon}_i$ is the function of all time delays $\eta_{1:M,i}$, which is calculated individually at each window. The integral combination that is closest to $\hat{\mathbf{P}}_i$ is taken as the solution to the period number, which is given by

$$p_{m,i}=\min_{p\in P_{m,i}}|p-\hat{p}_{m,i}|.\tag{48}$$

The optimal DOA estimate of the window-dominant source is eventually obtained by substituting the period numbers and the minimal time delays into (31). An iterative algorithm is presented to realize CF-WDSS, as shown in Algorithm 2. The weights of reliable delays are incrementally increased and the effects of unreliable delays are weakened with increasing iterations.

Algorithm 2 DOA estimation of the dominant source using CF-WDSS

```
Estimate the correlation matrix at a window using (8);
Perform eigenvalue decomposition and calculate $\eta_{m,i}$ using (26);
Estimate the period number using (47) and (48);
Set all weights as $w_{1:M,i} = 1$ and calculate $\hat{\boldsymbol{\gamma}}_i$ using (31);
repeat
    Let $\boldsymbol{\gamma}' = \hat{\boldsymbol{\gamma}}_i$ and calculate the new weights using (34).
    Calculate $\hat{\boldsymbol{\gamma}}_i$ with the new weights using (31);
until $(1 - \boldsymbol{\gamma}'^T\hat{\boldsymbol{\gamma}}_i < \epsilon)$
```

V. Implementation

The DOAs of speech sources are eventually obtained by using histogram analysis on the estimated DOAs of all window-dominant sources. Because azimuth discrimination is much more precise than elevation discrimination for an array with horizontal orientation, the speech sources are identified by using azimuths of dominant sources which are given by

$$\hat{\alpha}_i=\begin{cases}\arccos\Big(\hat{\gamma}_{1,i}/\sqrt{\hat{\gamma}^2_{1,i}+\hat{\gamma}^2_{2,i}}\Big) & if\ \hat{\gamma}_{2,i}\geq 0\\ 180^\circ+\arccos\Big(\hat{\gamma}_{1,i}/\sqrt{\hat{\gamma}^2_{1,i}+\hat{\gamma}^2_{2,i}}\Big) & if\ \hat{\gamma}_{2,i}<0.\end{cases}\tag{49}$$

The histogram with 1-degree resolution is constructed on the azimuths of all window-dominant sources. There are often some phantom peaks in the histogram. The Hanning smoothing window is helpful to remove phantom peaks. However, a window with too large size is likely to smooth out the real peaks. In the preliminary experiment, the 13-histogram-bin window makes a good tradeoff. The speech sources are identified by picking peaks in the histogram, as schematically illustrated in Fig. 7. Each source corresponds to a peak with occurrence greater than the threshold,

$$\Delta=O_{avg}+\mu(O_{max}-O_{avg}),\tag{50}$$

where O_{avg} and O_{max} denote the average of the smoothed occurrence and the maximum of the smoothed occurrence, respectively, and the coefficient μ $(0<\mu<1)$ is set empirically. A large value of the coefficient μ enables a conservative detection, which results in less false detections but more missed true detections. It is vice versa for a small value of the coefficient. The number of sources is determined by counting the distinct peaks.

The localization is conducted on every 15-frame segment. Since the five-frame window makes a good tradeoff, each speech segment is partitioned into three five-frame subsegments to mitigate the masking effect. Accordingly, the percentage of the masked bins is reduced from 48% to 17% at the cost of increasing the computational load. Each histogram is constructed on the azimuths from three 5-frame subsegments that are illustrated by the three overlapped dashed boxes in Fig. 6. The signal sampling rate, window size, frame length, and parameters about STFT were the same as that in Section III. Six low-frequency bins are disregarded ($h=7$). The algorithmic delay of the proposed methods is 0.24 seconds, which is much less than the algorithmic delay of most methods [29], [32]–[35], [37]–[46].

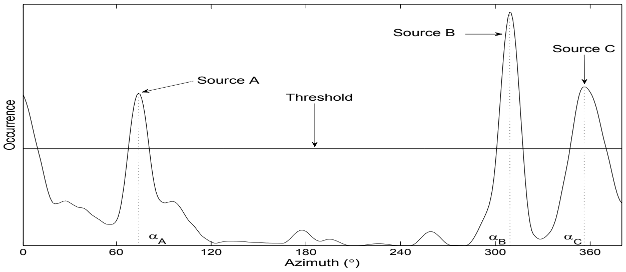

Fig. 7. Schematical illustration of the peak picking method.

VI. Evaluation

A. Performance

Evaluation was conducted in a simulated room with dimensions of $5 \times 8 \times 2.7$ m. An eight-element uniform circular array was placed with horizontal orientation at the center of the room. The radius of the circular array was 8 cm. The speech sources were located at a horizontal distance of 1.2 m from the array center. Their vertical heights were 0.96 m relative to the array plane. The scenarios were simulated using the image source method [50] - [52] to control reverberation times. The real room noise was artificially added to the simulated signals with various SNRs. For conventionally used continuous speech, it is very difficult to definitely count the actual number of sources in each short-term segment since the source signal may be weak in some times. For this reason, this evaluation was conducted on some specially-designed test samples that were made by mixing monosyllabic utterances with equal intensity. The presence of speech sources is ensured by the process of making the test samples, which is demonstrated by Fig. 8. The maximal intensity of each source utterance is aligned along the time. Each 15-frame test sample is artificially taken from the 7 preceding frames and 7 succeeding frames around the maximal-intensity frame. Note that the speech activity detection is out of the scope of this paper because it is an unknown factor in this evaluation. There are in total 138 test samples, 138×3 sources for the three-source evaluation, and 138×2 sources for the two-source evaluation.

GS-WDSS and CF-WDSS were compared with the broadband MUSIC and circular harmonics beamforming (CHB) [43] methods. CHB utilizes the conventional assumption on speech sparsity ($L = 1$). At each bin, the grid search is conducted to find the azimuth of the bin-dominant source. The speech sources are eventually identified by the azimuth histogram, which is the same as the proposed methods. The number of sources is determined by counting the significant peaks in the azimuth histogram for the proposed methods and CHB, and in the spatial power spectrum for MUSIC. It is worthwhile clarifying that, for MUSIC, the noise subspace is formed by using the known source number, which means that MUSIC cannot really determine the number of sources. MUSIC, GS-WDSS, and CHB perform grid search at 1-degree intervals. Since the horizontally-oriented array cannot provide precise discrimination of the elevations, the evaluations focused on the arrival azimuths.

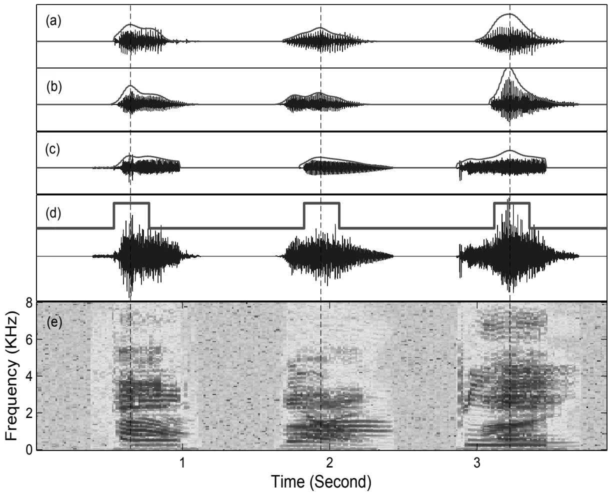

Fig. 8. Demonstration of making test samples: (a) Waveform and intensity of the first source; (b) Waveform and intensity of the second source; (c) Waveform and intensity of the third source; (d) Waveform of the mixing signal, where the dashed line denotes the 15-frame samples used for the test; (e) Spectrogram of the test signal. The vertical dotted lines denote the center of the test samples, which corresponds to the maximal intensity of each short-term utterance.

The evaluation for short-term speech source localization is more complicated than the conventional evaluation for long-term speech source localization. Because the sufficient frames that are provided by long-term utterances lead to few missed true detections or false detections (e.g. detected but non-existing sources), the conventional methods were usually evaluated by means of the localization accuracy such as the root mean squared error (RMSE) between the real and estimated DOAs. In short-term speech source localization, however, the missed detections and the false detections are frequently present. Besides the detection accuracy, the detection correctness should be considered in the evaluation. The output sources are classified into the correctly detected sources and the incorrectly detected sources. The detection is considered to be correct if the estimated azimuth deviates no more than 6° from the real azimuth of any source. The threshold for correctness can guarantee the performance of DOA-based enhancement and separation does not drop significantly according to our experience. The incorrect detections consist of the false detections and the inaccurate detections. The detection correctness is mainly accessed in terms of 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 (FDR) (i.e., the ratio of the number of incorrectly detected sources to the total number of sources). PDR and FDR jointly evaluate not only the detection correctness, but also the capability of counting the number of sources on all test segments. A large value of the coefficient μ results in small PDR and FDR, and vice versa for the small coefficient. The receiver operating characteristic (ROC) curve provides a complete description of the relationship between PDR and FDR under different coefficients.

The first experiment made use of ROC curves to evaluate the performance on the simulated data. Three speech sources were respectively located at the azimuth angles of 121°, 177°,

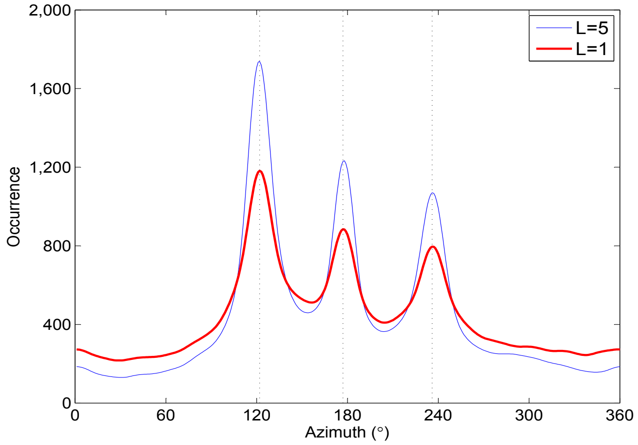

Fig. 10. The histogram of bin-wise azimuths that are estimated by CF-WDSS.

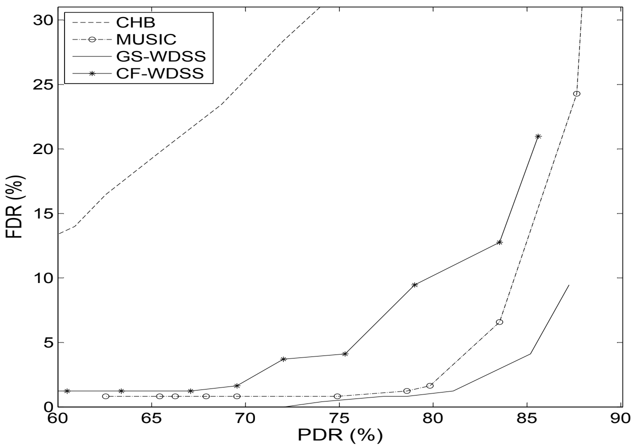

Fig. 12. ROC curves for the real data.

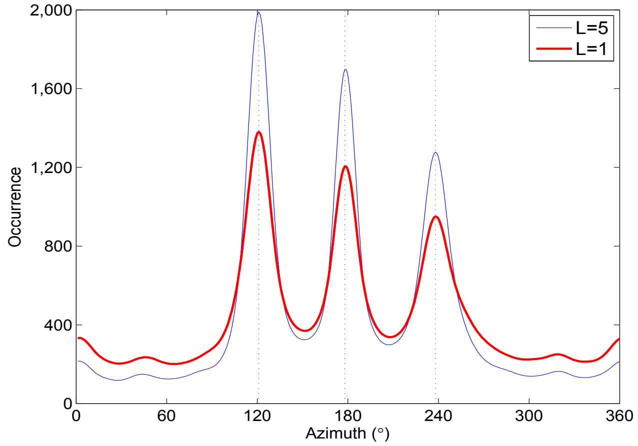

Fig. 11. The histogram of bin-wise azimuths that are estimated by GS-WDSS.

and $236°$. The ROC curves were obtained by tuning the coefficient μ from 0 to 0.4 with 10 equal spaces. Fig. 9 plots the ROC curves under nine conditions. The experiment shows that CHB does not work well on short-term segments, and MUSIC performs best under anechoic conditions. Especially in the least noisy and reverberant condition of (20 dB, 0 ms), MUSIC can correctly detect all sources without false detections. However, the performance of MUSIC significantly degrades with increasing reverberation times. In reverberant conditions, GS-WDSS outperforms the remaining methods. The same conclusion was obtained with the correctness threshold ranging from 6 to 10 degrees in the preliminary experiment.

At the most adverse condition ($SNR = 10$ dB, $T60 = 400$ ms), a comparison is made to illustrate the advantage of using a longer window. The smoothed histograms are constructed on the bin-wise azimuths that are estimated on 138 test samples, as shown in Figs. 10 and 11. The comparison confirms the superiority of the generalized assumption.

The second experiment was conducted in a real environment. It was used to confirm the results of the simulated experiments. The performance was evaluated in a room with a reverberation time of 250 ms and SNR of about 15 dB. There were 80 short-term utterances and 80×3 speech sources used for evaluation. The speakers try to synchronize with each other and pronounce short utterances with equal intensity. The remaining experimental setup was the same as that of the first simulated experiment. The ROC curves are shown in Fig. 12. Although we try to make the same acoustic condition as the simulated condition, there are still some differences between them. Nevertheless, the conclusion of the real experiment is similar to that of the simulated experiment.

The histogram was utilized to intuitively compare the performance for two given sources with different azimuth spacings. We simulated an moderately adverse environment with the reverberation time of 250 ms and the SNR of 10 dB. Two speakers were respectively located at azimuths with spacing of $14°$, $24°$, and $34°$. The coefficient μ was set to 0.2 for all methods. The remaining setup is the same as that in the aforementioned experiments. Fig. 13 plots the histograms for the three spacings of azimuth angles, wherein the vertical dotted lines denote the real azimuths. This figure shows that MUSIC is incapable of discriminating the spatially-close sources. The three sparsity-based methods have a significant advantage over MUSIC in the spatial resolution. For two widely spaced sources, the proposed methods still perform better than CHB and MUSIC. Actually, CHB produces more false detections than the remaining methods, which are plotted outside the range of the azimuth in Fig. 13.

The localization accuracy was evaluated on both the simulated and the real data. For a given speech source, RMSE is averaged over the 69 most accurate detections of the simulated data, and the 40 most accurate detections of the real data, wherein the false detections are excluded from evaluation because they are meaningless for the localization accuracy. The number of the accurate detections varies over methods and acoustic conditions. The used number can guarantee that the false detections are excluded from the evaluation for all conditions and all methods. Afterwards, RMSE for each method is averaged over three sources. Table. I presents RMSE for four methods. The result shows that MUSIC performs the most accurately in anechoic conditions while GS-WDSS performs the most accurately in the remaining conditions. The localization accuracy of CF-WDSS is even higher than that of

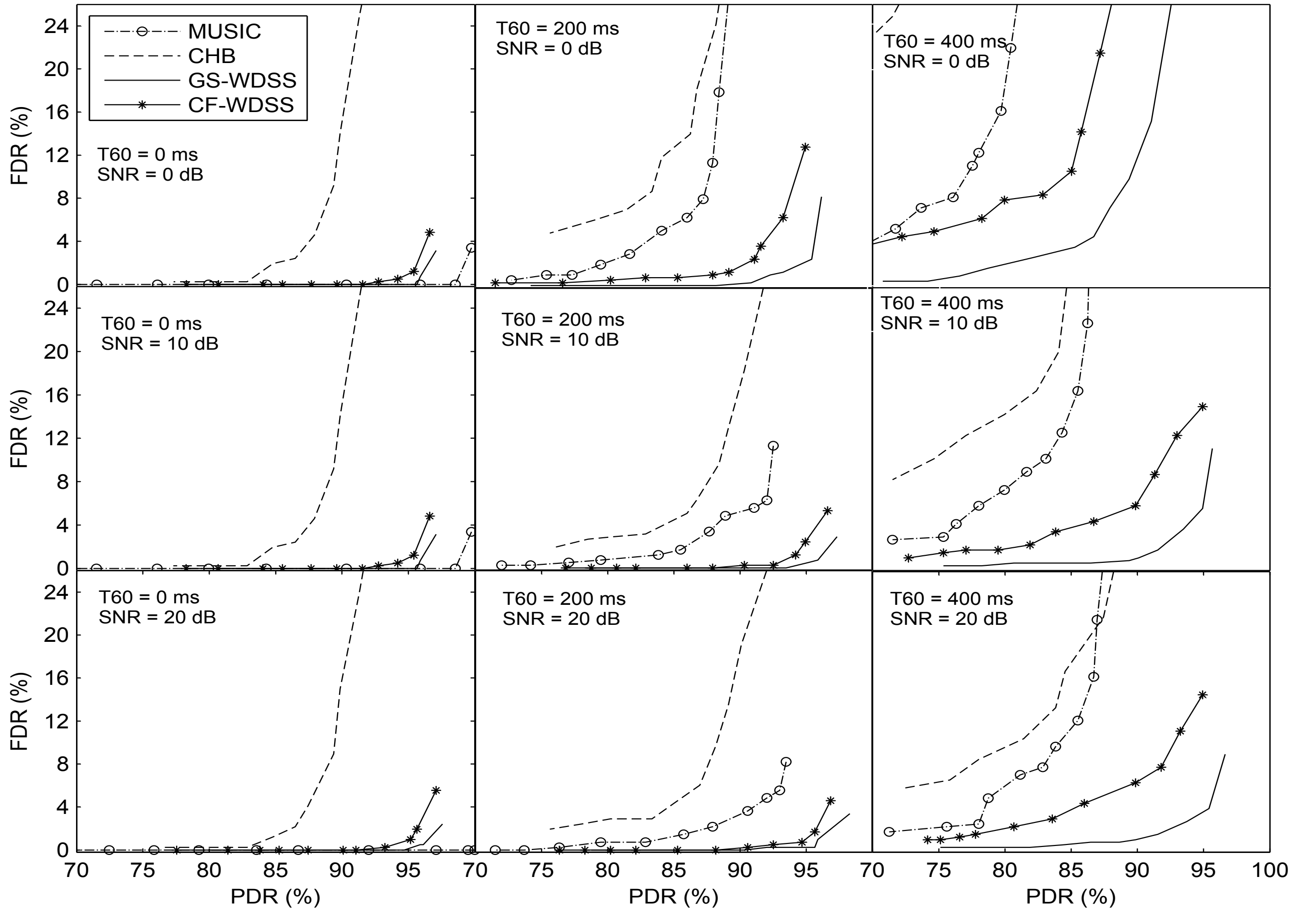

Fig. 9. ROC curves for the simulated data set under various conditions.

TABLE I
RMSE (°) FOR DIFFERENT CONDITIONS.

Method	$T60 = 0$ ms			$T60 = 200$ ms			$T60 = 400$ ms			Real data
	0 dB	10 dB	20 dB	0 dB	10 dB	20 dB	0 dB	10 dB	20 dB	
MUSIC	**0**	**0**	**0**	0.52	0.38	0.36	1.39	1.19	1.18	0.56
CHB	0.77	0.53	0.45	1.25	0.93	0.90	1.62	1.38	1.33	1.30
GS-WDSS	**0**	**0**	**0**	**0.12**	**0.07**	**0.02**	**0.48**	**0.47**	**0.45**	**0.34**
CF-WDSS	0.29	0.21	0.19	0.25	0.19	0.18	0.53	0.35	0.33	0.86

MUSIC in the condition of $T60 = 400$ ms.

Lastly, we investigated the iteration for CF-WDSS in Algorithm 1. For a given iteration number, the convergence error is averaged over all times, all frequencies, and all conditions on the first simulated data. The convergence error is generally reduced with increasing iterations. On average, the error is reduced by 3° within three iterations, and by a little bit afterwards. For the sake of computational efficiency, the iteration is conducted at most three times in CF-WDSS.

B. Computational load

The computational loads of four methods were compared. The computational loads were calculated by counting the number of basic operations in the source code. For MUSIC, the orthogonality test is conducted on $\mathcal{C}(\alpha) \times \mathcal{C}(\beta)$ grids to search for the local minima. The noise subspace consists of $K - D$ eigenvectors. Therefore, the computational load mainly consists of $\mathcal{C}(\alpha) \times \mathcal{C}(\beta) \times \mathcal{C}(i) \times K \times (K - D)$ complex multiplications and complex additions. GS-WDSS only uses the principal eigenvector, whereas the grid search is repeated three times for each localization to mitigate the masking effect. Its computational load is about $3/(K - D)$ of MUSIC. CHB tests the harmonic beamforming on all potential azimuths, all frames, and all frequencies. The elevation is disregarded in the test. Its computational load consists of $15 \times \mathcal{C}(i) \times \mathcal{C}(\alpha)$ bin-wise beamformings. For CF-WDSS, an iterative procedure is repeated no more than three times. The DOA estimation is conducted at three subsegments and

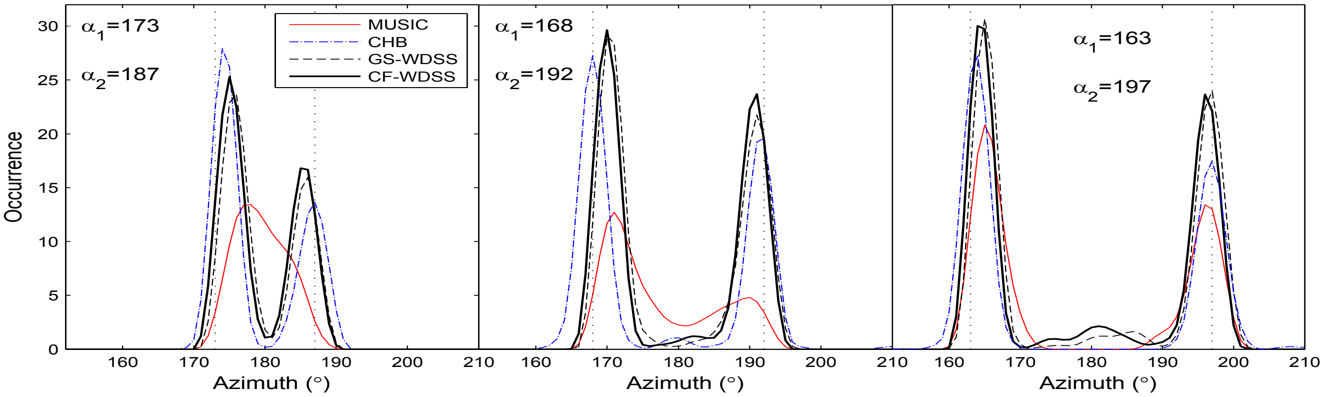

Fig. 13. Azimuth histograms of two sources with various azimuth spacings.

TABLE II
COMPUTATIONAL LOAD COMPARISON

Method	MUSIC	CHB	GS-WDSS	CF-WDSS
Relative load	55.4	2.5	34.0	1
Per. of eigen decomp.	0.45%	0%	2.2%	73.2%
Per. of bin-wise proces.	99.3%	97.2%	97.3%	5.6%

$\mathcal{C}(i)$ frequencies. Therefore, the closed-form solution to the dominant source DOA is calculated at most $3 \times 3 \times \mathcal{C}(i)$ times. The eigen-decomposition is another factor that results in computational load for MUSIC, GS-WDSS, and CF-WDSS. In each localization, the decomposition is conducted $\mathcal{C}(i)$ times for MUSIC, and $3 \times \mathcal{C}(i)$ times for CF-WDSS and GS-WDSS.

According to the computing process of the four methods, this paper utilizes the relative computational load, the load percentage of the eigen-decomposition, and the percentage of bin-wise processing, to evaluate the computational efficiency. Table II compares the computational loads, where the relative load is taken with regard to CF-WDSS. The eigenvalue decomposition is the major computational load for CF-WDSS while the loads of the remaining methods are mainly contributed by the grid search. Given the parameter settings in this paper, the computational load of decomposition is much smaller than that of grid search, and the computational efficiency is sorted in the descending order, CF-WDSS>CHB>GS-WDSS>MUSIC. It should be noticed that MUSIC and GS-WDSS require extra memory to storage the steering vectors on all potential directions ($\mathcal{C}(\alpha) \times \mathcal{C}(\beta) \times \mathcal{C}(i)$ steering vectors and 118-megabyte memory for the experimental settings) whereas CF-WDSS does not.

C. Discussion

The performance of four localization methods are evaluated using short-term speech segments with duration of 0.24 seconds. Since insufficient frames are provided to four methods in each localization, the robustness are highlighted in the experiments. The proposed methods achieve the best performance under heavy reverberation while they are inferior to MUSIC in anechoic environments. The performance of the proposed methods is under the influence of two contradictory points. One is the DOA estimation of the window-dominant source, which is highly correlated with the similarity between the steering vector and the principal eigenvector. Enlarging the window size is helpful to improve the performance of the dominant source DOA estimation. The other point is the masking effect on weak speech sources. Enlarging the size will enhance the masking effect, which leads to less windows to be dominated by the weak speech sources. As a result, it is difficult to identify weak sources in the histogram. Even in an ideal environment without any interference, the proposed methods may not correctly detect all sources due to the masking effect. The experiments in Section III showed that the five-frame window makes a good tradeoff between the two points.

CHB utilizes the conventional sparsity assumption in contrast to the generalized assumption adopted in the proposed methods. Although the masking effect in CHB is much smaller than that in the proposed methods, CHB does not perform well on short-term speech source localization. Its poor performance implies that CHB relies heavily on the sufficient frames to achieve the robust localization. In fact, CHB was reported to work fairly well on long-term utterances with duration of 4 seconds [43]. The experiments demonstrate that the proposed methods substantially outperform CHB on short-term speech source localization, confirming the superiority of the generalized assumption.

MUSIC does not utilize the speech sparsity assumption. Its performance is best in anechoic environments, but significantly deteriorated in reverberant environments. MUSIC has no masking effect on weak speech sources, and therefore, the sources can be well detected in less reverberant environments. MUSIC makes use of the noise subspace as opposed to the principal eigenvector adopted in the proposed methods. The proposed methods have three advantages over MUSIC. The first is the computational efficiency, as shown in Table II. The second is the robustness over reverberation. MUSIC can estimate not only the DOA of the dominant source, but also the DOAs of the masked sources at each frequency bin. However, the estimation for the masked sources usually suffers from reverberation since the masked sources generally associate with weak speech components. On the contrary, the principal eigenvector generally associates with the direct-path signal, which is not likely to be affected by the reverberation. The last advantage is the capability of counting the number of speech sources. MUSIC cannot count the number of speech sources. It should be noted that MUSIC does not perform as well in reality as in the experiments since the number of sources is difficult to always estimate accurately.

VII. Conclusions

Two WDSS methods were developed to localize multiple speech sources using short-term segments. By tuning the window size for the generalized sparsity assumption, a tradeoff is made between the masking effect and the robustness in the DOA estimation for dominant sources. The proposed methods not only have the advantage in robustness, but also can count the number of speech sources. Although CF-WDSS does not perform as well as GS-WDSS, the former is very computationally efficient to localize speech sources. The proposed methods are valuable for real-time speech source localization because of the small latency. The generalized sparsity assumption makes promises to improve speech source localization in adverse environments. Besides time-adjacent bins, frequency-adjacent bins can be considered for the generalized assumption, which will be addressed in our future work.

Acknowledgment

We would like to thank the anonymous reviewers for their constructive suggestions. We also thank Prof. Qixiang Ye for his helpful comments to the technique writing. This study is supported by National Program on Key Basic Research Project (2013CB329302), National Natural Science Foundation of China (Nos. 61671442, 61271426, 11461141004, 91120001) and the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant Nos. XDA06030100, XDA06030500).

References

[1] C. Busso, S. Hernanz, C. Chu, S. Kwon, S. Lee, P. Georgiou, I. Cohen, and S. Narayanan, "Smart room: participant and speaker localization and identification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Philadelphia, USA, 2005, pp. 1117–1120.

[2] J. Nikunen and T. Virtanen, "Direction of arrival based spatial covariance model for blind sound source separation," *IEEE/ACM Trans. on Audio, Speech, and Language Process.*, vol. 22, no. 3, pp. 727–739, 2014.

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

[4] Y. Huang, J. Benesty, and G. W. Elko, "Passive acoustic source localization for video camera steering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Istanbul, Turkey, 2000, pp. 909–912.

[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 Language Process.*, vol. 21, no. 4, pp. 828–841, 2013.

[6] Z. Yu and Y. Nakamura, "Smart meeting systems: A survey of state-of-the-art and open issues," *ACM Computing Surveys*, vol. 42, no. 2, pp. 926–934, 2010.

[7] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, 1986.

[8] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 7, pp. 984–995, 1989.

[9] G. Su and M. Morf, "Signal subspace approach for multiple wideband emitter location," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-31, no. 7, pp. 1502–1522, 1983.

[10] W. Zeng and X. Li, "High-resolution multiple wideband and nonstationary source localization with unknown number of sources," *IEEE Trans. Signal Process.*, vol. 58, no. 6, pp. 3125–3136, 2010.

[11] F. Asano, H. Asoh, and K. Nakadai, "Sound source localization using joint Bayesian estimation with a hierarchical noise model," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 21, no. 9, pp. 1953–1965, 2013.

[12] H. Wang and M. Kaveh,, "Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 4, pp. 823–831, 1985.

[13] S. Talagala, W. Zhang, and D. Abhayapala, "Broadband DOA estimation using sensor arrays on complex-shaped rigid bodies," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 21, no. 8, pp. 1573–1584, 2013.

[14] D. Swingler and J. Krolik, "Source location bias in the coherently focused high-resolution broadband beamformer," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 1, pp. 143–145, 1989.

[15] T. Lee, "Efficient wide-band source localization using beamforming invariance technique," *IEEE Trans. on Signal Process.*, vol. 42, no. 10, pp. 1376–1387, 1994.

[16] E. Claudio and R. Parisi, "WAVES: Weighted average of signal subspaces for robust wideband direction finding," *IEEE Trans. on Signal Process.*, vol. 49, no. 10, pp. 2179–2190, 2001.

[17] 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.

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

[19] J. Dmochowski, J. Benesty, and S. Affes, "Direction of arrival estimation using the parameterized spatial correlation matrix," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 15, no. 4, pp. 1327–1339, 2007.

[20] Y. Wu, A. Leshem, J. Jensen, and G. Liao, "Joint pitch and DOA estimation using the ESPRIT method," *IEEE/ACM Trans. on Audio, Speech, and Language Process.*, vol. 23, no. 1, pp. 32–45, 2015.

[21] J. Zhang, M. Christensen, S. Jensen, and M. Moonen, "Joint DOA and multi-pitch estimation based on subspace techniques," *EURASIP J. Adv. Signal Process.*, vol. 2012, no. 1, pp. 1–11, 2012.

[22] F. Belloni, A. Richter, and V. Koivunen, "DoA estimation via manifold separation for arbitrary array structures," *IEEE Trans. Signal Process.*, vol. 55, no. 10, pp. 4800–4810, 2007.

[23] C. Mathews and M. Zoltowski, "Eigenstructure techniques for 2-D angle estimation with uniform circular arrays," *IEEE Trans. Signal Process.*, vol. 42, no. 9, pp. 2395–2407, 1994.

[24] M. Costa, A. Richter, and V. Koivunen, "DoA and polarization estimation for arbitrary array configurations," *IEEE Trans. Signal Process.*, vol. 60, no. 5, pp. 2330–2343, 2012.

[25] D. Zotkin and R. Duraiswami, "Accelerated speech source localization via a hierarchical search of steered response power," *IEEE Trans. on Audio, Speech Process.*, vol. 12, no. 5, pp. 499–508, 2004.

[26] J. Dmochowski, J. Benesty, and S. Affes, "A generalized steered response power method for computationally viable source localization," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 15, no. 8, pp. 2510–2526, 2007.

[27] L. Nunes, W. Martins, M. Lima, L. Biscainho, M. Costa, M. Goncalves, A. Said, and L. Bowonee, "A steered-response power algorithm employing hierarchical search for acoustic source localization using microphone arrays," *IEEE Trans. on Signal Process.*, vol. 62, no. 19, pp. 5171–5183, 2014.

[28] D. Yook, T. Lee, and Y. Cho, "Fast Sound Source Localization Using Two-Level Search Space Clustering," *IEEE Trans. on Cybernetics*, vol. 46, no. 1, pp. 20–26, 2016.

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

[30] P. Bofill and M. Zibulevsky, "Blind separation of more sources than mixtures using sparsity of their short-time Fourier transform," in *Proc. Int. Workshop Independent Component Anal. Blind Signal Separation*, Helsinki, Finland, 2000, pp. 87–92.

[31] 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.

[32] S. Brandstein and F. Silverman, "A practical methodology for speech source localization with microphone arrays," *Computer Speech and Language*, vol. 11, no. 2, pp. 91–126, 1997.

[33] C. Liu, B. Wheeler, W. Brien, 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.

[34] M. Mandel, D. Ellis, and T. Jebara, "EM algorithm for localizing multiple sound sources in reverberant environments," in *Advances in Neural Information Processing Systems*, B. Schökopf, J. Platt, and T. Hoffman, Eds. Cambridge, MA: MIT Press, 2007, pp. 953–960.

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

[36] X. Zhong and J. Hopgood, "A time-frequency masking based random finite set particle filtering method for multiple acoustic source detection and tracking," *IEEE/ACM Trans. on Audio, Speech, and Language Process.*, vol. 23, no. 12, pp. 2356–2370, 2015.

[37] 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.

[38] W. Xue, W. Liu, and S. Liang, "Noise robust direction of arrival estimation for speech source with weighted bispectrum spatial correlation matrix," *IEEE J. of Selected Topics in Signal Process.*, vol. 9, no. 5, pp. 837–851, 2015.

[39] M. Kühne, 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.

[40] 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.

[41] S. Araki, H. Sawada, R. Mukai, and S. Makino, "DOA estimation for multiple sparse sources with normalized observation vector clustering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Toulouse, France, 2006, pp. 33–36.

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

[43] 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.

[44] Q. Shen, W. Liu, W. Cui, S. Wu, Y. Zhang, and M. Amin, "Low-complexity direction-of-arrival estimation based on wideband co-prime arrays," *IEEE/ACM Trans. on Audio, Speech, and Language Process.*, vol. 23, no. 9, pp. 1445–1456, 2015.

[45] O. Nadiri and B. Rafaely, "Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-path dominance test," *IEEE/ACM Trans. on Audio, Speech, and Language Process.*, vol. 22, no. 10, pp. 1494–1505, 2014.

[46] S. Mohan, M. Lockwood, M. Kramer, and D. Jones, "Localization of multiple acoustic sources with small arrays using a coherence test," *J. Acoust. Soc. Am.*, vol. 123, no. 4, pp. 2136–2147, 2008.

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

[48] D. Brungarta, B. Simpson, M. Ericson, and K. Scott, "Informational and energetic masking effects in the perception of multiple simultaneous talkers," *J. Acoust. Soc. Am.*, vol. 110, no. 5, pp. 2527–2538, 2001.

[49] J. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database,", *Nat. Inst. Standards Technol. (NIST)*, Gaithersburg, MD, prototype as of Dec. 1988.

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

[51] E. Lehmann and A. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *J. Acoust. Soc. Am.*, vol. 124, no. 1, pp. 269–277, 2008.

[52] http://www.eric-lehmann.com/ism_code.html

[53] 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.

PLACE PHOTO HERE

Dongwen Ying received his B.E. degree in 1998 and his M.E. degree in 2000, respectively, from Harbin Institute of Technology. He earned his Ph.D. from the School of Information Science, Japan Advanced Institute of Science and Technology (JAIST), in 2007. He is currently working at ThinkIT Lab, Institute of Acoustics, Chinese Academy of Sciences as a researcher. His major interests include speech enhancement, voice activity detection, sound source location, speech separation, and robust speech recognition.

PLACE PHOTO HERE

Ruohua Zhou was born in 1972. He received Bachelor Degree in Electronics Engineering Department of Beijing Institute of Technology, China, in 1994; He received Master degree of engineering in microelectronics and semiconductor devices from Chinese Academy of Sciences, Microelectronics R&D Center, Beijing, in 1997 and PhD degree on audio signal processing from the Signal Processing Institute of Swiss Federal Institute of Technology, Lausanne, in 2006. Currently, He is a professor in the Institute of Acoustics (IOA) of the Chinese Academy Sciences (CAS).

PLACE PHOTO HERE

Junfeng Li received the Ph.D. degree in Information Science from Japan Advanced Institute of Science and Technology (JAIST) in March 2006. From April 2006, he was a post-doctoral research fellow at Research Institute of Electrical Communication (RIEC), Tohoku University. From April 2007 to July 2010, he was an Assistant Professor in School of Information Science, JAIST. Since August 2010, he has been a Professor in Institute of Acoustics, Chinese Academy of Sciences. His research interests include psychoacoustics, acoustic signal processing and 3D audio technology. Dr. Li received the Best Student Award in Engineering Acoustics First Prize from the Acoustical Society of America in 2006, and the Best Paper Award from JCA2007 in 2007, and the Itakura Award from the Acoustical Society of Japan in 2012. Dr. Li is now serving as Subject Editor for Speech Communication and Editor for IEICE Trans. on Fundamentals of Electronics, Communication and Computer Sciences.

PLACE PHOTO HERE

Yonghong Yan received his B.E. from Tsinghua University in 1990 and his Ph.D. from Oregon Graduate Institute (OGI). He worked in OGI as Assistant Professor (1995), Associate Professor (1998) and Associate Director (1997) of Center for Spoken Language Understanding. He worked in Intel from 1998-2001, chaired Human Computer Interface Research Council, worked as Principal Engineer of Microprocessor Research Lab and Director of Intel China Research Center. Now he is a professor and director of ThinkIT Lab. His research interests include speech processing and recognition, language/speaker recognition and human computer interface. He has published more than 100 papers and holds 40 patents.