Noise Robust Direction of Arrival Estimation for Speech Source With Weighted Bispectrum Spatial Correlation Matrix

Wei Xue, Wenju Liu, and Shan Liang

***Abstract*—One big challenge to the robust direction of arrival (DOA) estimation for the speech source is the environmental noise. In practical conditions, the noise can be undirected or emitted from a pointed source. In order to improve the reliability of DOA estimation in various adverse noisy conditions, we propose a novel DOA estimation method in this paper, and what lies in the core in the method is the "Weighted Bispectrum Spatial Correlation Matrix (WBSCM)." The bispectrum is a kind of higher order statistics (HOS) of a signal, and the WBSCM reflects the spatial correlation of the bispectrum phase differences (BPD) between different microphones. As the HOS of the Gaussian signal is theoretically zero, by formulating in the bispectrum domain, the proposed method has an inherent advantage against the Gaussian noise. Moreover, the BPD, which is embedded in the WBSCM, contains the redundant information related to the DOA of the speech source. This redundancy helps to improve the robustness in non-Gaussian noise conditions, especially for the directional interference scenarios. In addition, the WBSCM enables bispectrum weighting to select the speech units in the bispectrum, in order to highlight the effect of these units in the DOA estimation. Similar to the signal-to-noise estimation, a decision-directed method is proposed to compute the bispectrum weights. Finally, a new DOA estimator is proposed, which is based on the eigenvalue analysis of the WBSCM. We conduct experiments under various kinds of noisy environments, and the experimental results demonstrate the effectiveness of proposed method.**

***Index Terms*—Direction of arrival (DOA) estimation, microphone array signal processing, high order statistics, bispectrum.**

I. Introduction

A MAJOR functionality of the microphone array is to estimate the direction of arrival (DOA) of the sound source. In many speech processing systems, such as the hands-free devices, video conference systems, and interactive robots, the knowledge of DOA of the speech source is of great interest [1]–[4]. In practical scenarios, the noise which exists in the surrounding environment always makes the DOA estimation a much complicated problem.

Manuscript received July 14, 2014; revised January 04, 2015; accepted March 06, 2015. Date of publication March 25, 2015; date of current version July 14, 2015. This work was supported in part by the China National Nature Science Foundation under Grants 91120303, 61273267, 90820011, 90820303, and 61403370. The guest editor coordinating the review of this manuscript and approving it for publication was Prof. Ramani Duraiswami.

The authors are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: wxuenlpr.ia.ac.cn; lwjnlpr.ia.ac.cn; sliang@nlpr.ia.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2015.2416686

Over the past few decades, many kinds of DOA estimation methods for the speech source have been developed. In some fields of the narrow-band signal processing, such as radar and underwater sonar, estimating the DOAs of narrow-band sources has been a crucial problem for a long history, and some high-resolution spectral methods have been proposed, e.g., the MUSIC algorithm [5], the Capon's minimum variance (MV) algorithm [6]. High-resolution spectral methods for the broadband speech are generally developed from these narrow-band approaches. These methods usually decompose the broadband signal into several narrow-band signals [7], [8], or utilize a "focusing matrix" to transform the signal of all narrow-bands into one reference narrow band [9], [10]. A more straightforward approach is proposed in [11], where a time-domain formulation is derived. Another category of DOA estimation methods are the steering response power (SRP) based methods [12]–[14], which utilize a beamformer to steer over all potential directions, then determine the DOA of speech source by searching for the direction with the largest SRP. Compared with above two categories of methods, the time delay estimation (TDE) based methods have received much more attention. TDE based methods are two-step approaches, which estimate the time delays between multiple microphone pairs firstly, and then compute the DOA according to the time delays and the array geometry [15], [16]. For pairwise TDE, the most widely used methods are the generalized cross-correlation (GCC) family of methods [17]. However, they can only be used pairwise, and often suffer from performance degradation in highly noisy conditions [18]. Some authors propose to exploit the spectral properties of the speech signal for TDE, for example, the harmonic structure of reverberated speech is considered in [19], and the features of the excitation source of voiced speech are utilized in [20]. Several other methods suggest employing more microphones to improve the performance, e.g., the spatial linear prediction (SLP) method and the multichannel cross-correlation coefficient (MCCC) algorithm [21]–[23].

Although many approaches have been proposed, most of them assume that the additive noises at different microphones are white Gaussian signals and uncorrelated pairwise. Unfortunately, the assumption does not always hold in practical scenarios, as the noise at one individual microphone is not necessarily white Gaussian, and the noises at different microphones may not be spatially uncorrelated. When directional interferences e.g., computer fans and air conditioners exist, the noises at different microphones are spatially correlated. Due to the model inaccuracy, traditional methods tend to simply

estimate the DOA of the stronger signal [24], thus they can not distinguish between a speaker and a directional interference. Since the speech signal is non-stationary, even if the average signal to noise ratio (SNR) is higher than 0 dB, the speech may not be stronger than the interference in all time frames. This makes the DOA estimator can not stick to the DOA of the speech source in interference-existing scenarios.

Several methods have been developed to cope with the interference. In [25], the authors first estimate the DOAs of multiple sound sources, then use Gaussian mixture models (GMMs) to identify whether the beamformed signal from each candidate DOA is speech. As more than one frame is needed for identification, and the identification is performed after pre-estimating the multiple DOAs, this approach can not be used for on-line frame-level estimation. A different approach is proposed in [24], where the pairwise time difference of arrival (TDOA) is estimated by finding the peak location of the time-domain "Acoustic Transfer Function's Ratio," which is computed in the frequency domain by exploiting the speech signal's quasi-stationary and interference signal's stationary. However, the extension from the pairwise case to the multichannel case for estimating the DOA of the far field source seems not straightforward. In [26], a interference robust DOA estimator which is applicable for the multichannel case is presented. The DOA estimator is formulated in the frequency domain, and it adopts frequency weighting to select only speech frequency bands so as to improve the robustness against the interference. Clearly, among the whole set of frequency bands, only a few of them are speech ones, which contain the cues related to the DOA of speech source. Therefore, for the frequency domain formulated approaches, such as [24] and [26], if the noise signal has a flat frequency distribution, the DOA cues in speech frequency bands will be more likely polluted by the noise, making the DOA estimation less robust.

In this paper, we propose a method which formulates the DOA estimation problem in the "bispectrum" domain. Analogous to the second order spectrum—power spectrum, the bispectrum is the third order spectrum of a signal, which is a kind of "higher-order statistics" (HOS). In fact, there exists a promising property that the HOS of the Gaussian noise is theoretically zero. Therefore, several HOS based methods have been developed aiming to improve the performance in Gaussian noise conditions [27]–[31]. However, these methods are generally proposed for narrow-band signals, and it is always much time-consuming to decompose the speech signal into narrow-band signals, and then apply these approaches. More importantly, these methods are not capable to deal with non-Gaussian directional interferences whose HOS are non-zero.

Here, we formulate the algorithm in the bispectrum domain from a novel perspective, and make the resulting DOA estimator robust to both the Gaussian and non-Gaussian noises, and can be straightforwardly applied for broadband speech and multichannel cases. Compared with the frequency domain formulation, besides the immunity to Gaussian noise, the proposed method exploits another theoretical property that the speech DOA cues are expressed redundantly in the "bispectrum phase difference (BPD)" between two microphone signals. This redundancy has two advantages. Firstly, it means that even the DOA cue of the speech source is polluted in one bispectrum unit, it may be re-found in other unpolluted units. By contrast, for the frequency domain representation, once the cue in one speech band is polluted by the noise, it cannot be found elsewhere. This advantage makes it possible to improve the DOA estimation performance, especially in non-Gaussian directional noises. Secondly, as the speech DOA cues are repeatedly expressed, we can check whether the candidate DOA matches the real one by utilizing more than only one set of DOA cues, thus it can be viewed that the redundancy brings more observations for DOA estimation.

The core of the new DOA estimation method is the "Weighted Spatial Bispectrum Correlation Matrix (WBSCM)." The WBSCM contains the spatial correlation information of BPDs between different microphones. It provides a compact framework to exploit both the immunity of bispectrum against Gaussian noise and the redundancy of speech DOA cues expressed in the BPD for robust DOA estimation. In addition, the WBSCM enables bispectrum weighting to select the speech units in the bispectrum, in order to highlight the effect of speech bispectrum units in the DOA estimation. Similar to the SNR estimation, a decision-directed method is proposed to compute the bispectrum weights. The WBSCM is also a function of a hypothesized DOA, and has an interesting property only when the hypothesized DOA equals to the true one. A new DOA estimator is then proposed, which is based on the eigenvalue analysis of the WBSCM. We conduct experiments under various kinds of noisy environments, and the experimental results demonstrate the effectiveness of proposed method.

This work is an extension of our previous work in [32], [33]. The outline of this work is as follows. Section II presents the signal model and some assumptions. In Section III we present some basic concepts on the bispectrum and analyze the BPD between a pair of microphone signals. The details of the WBSCM including the calculation of the bispectrum weights and the formulation of WBSCM will be presented in Section IV. Then in Section V, we introduce a new DOA estimator. The experimental results are given in Section VI. Finally, we conclude this paper in Section VII.

II. Signal Model

As shown in Fig. 1, suppose that there exists one speech source and several interferences in the far field, and the interferences are all uncorrelated with the speech source. We use a microphone array consisting of M elements to collect the sound signals. For each sound source, the signal propagates from the source to microphone as a plane wave, and the sound level falls as a function of the distance between the source and microphone [34].

Let us choose the first microphone as the reference point, then the signal received by the mth microphone $(m = 1, \ldots, M)$ at time k can be simply expressed as:

$$\begin{aligned} y_m(k) &= \alpha_m s(k - t_0 - \tau_{m1}) + v_m(k) + n_m(k) \\ &= \alpha_m s_m(k) + v_m(k) + n_m(k) \end{aligned} \tag{1}$$

where $s_m(k) = s(k - t_0 - \tau_{m1})$ is the unattenuated speech signal received by the mth microphone, which is a delayed version of the signal at the location of the speech source $s(k)$, with t_0 as the propagation time from the speech source to the

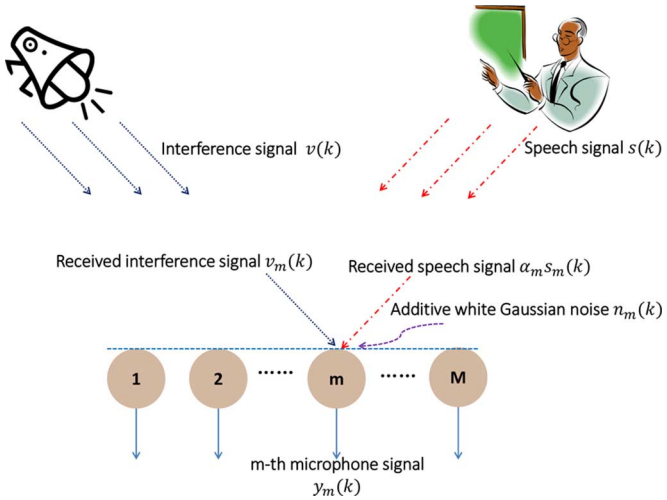


Fig. 1. Illustration of the signal model. $s_m(k)$ is the unattenuated speech signal received by the mth microphone, and α_m is the attenuation factor. Other notations are described inside the figure.

first microphone, and τ_{m1} as the relative delay between the mth and first microphone. α_m denotes the attenuation factor which ranges in $[0, 1]$. As only the speech signal which is related to the DOA of speech source is of our interest, we ignore the details of the interference signal received by the mth microphone, and simply represent the interference signal as $v_m(k)$. In addition, the $n_m(k)$ stands for the additive zero-mean white Gaussian noise.

Obviously, $\tau_{11} = 0$, then $s_1(k) = s(k - t_0)$, and $s_m(k) = s_1(k - \tau_{m1})$. Consequently, $y_m(k)$ can be rewritten as:

$$\begin{aligned} y_m(k) &= \alpha_m s_1(k - \tau_{m1}) + v_m(k) + n_m(k) \\ &= \alpha_m s_1(k - f_m(\varphi)) + v_m(k) + n_m(k). \end{aligned} \tag{2}$$

The time delay τ_{m1} is closely related to the geometry of the microphone array and real speech DOA φ. If the array geometry is fixed, τ_{m1} depends only on φ, then we use $f_m(\varphi)$ to denote the dependency. The mathematical formulation of $f_m(\varphi)$ can be well defined by geometrical computations. For example, a typical type of microphone array is the "uniform linear microphone array (ULA)," in which the array elements are equispaced, and in such case, we have:

$$\tau_{m1} = f_m(\varphi) = (m-1)\frac{\sin(\varphi) f_s d}{c}, m = 1, 2, \ldots, M, \tag{3}$$

where c is the speed of sound in the air, f_s is the sampling rate, and d is the spacing between two adjacent microphones.

III. Phase Difference in the Bispectrum Domain

A. Definitions and Properties of Bispectrum

In signal processing, one common way to describe the statistical properties of stochastic processes is to use the measures of second-order statistics, which generally include the auto-correlation, cross-correlation, and the corresponding power spectrum and cross-power spectrum. While the second-order statistics are widely used in various fields of signal processing, these measures only provide partial descriptions of the statistical properties of stochastic processes [35]. Therefore, the principles of correlations and power spectra have been extended to orders greater than two, and the concepts of HOS of stochastic processes are then introduced [36]–[38]. HOS generally include the higher-order moment, higher-order cumulant and the corresponding higher-order spectrum of stochastic processes. The "bispectrum," which is defined in the order of three, is the simplest higher-order spectrum. In the literature, for stationary stochastic signals, analog to the definition of power spectrum, the bispectrum is defined as the 2-D Discrete Fourier Transform (DFT) of the third-order cumulant of these stochastic signals [39].

Now let us consider the bispectrum[1] of three zero-mean stationary stochastic signals, which are denoted as $a(k), b(k)$ and $c(k)$. For zero-mean processes, the third-order cumulant is identical to the third-order moment, then the bispectrum $B_{abc}(\Omega_1, \Omega_2)$ of $a(k), b(k), c(k)$ is defined with the following expression:

$$B_{abc}(\Omega_1, \Omega_2) \stackrel{\text{def}}{=} \sum_{\rho_1=-\infty}^{\infty} \sum_{\rho_2=-\infty}^{\infty} \mathcal{R}_{abc}(\rho_1, \rho_2) e^{-j(\Omega_1 \rho_1 + \Omega_2 \rho_2)}, \tag{4}$$

where Ω_1 and Ω_2 are angular bi-frequency variables, $j = \sqrt{-1}$ is the imaginary unit, and $\mathcal{R}_{abc}(\rho_1, \rho_2)$ is the third-order moment of $a(k), b(k), c(k)$, which is defined depending on two independent lags ρ_1 and ρ_2:

$$\mathcal{R}_{abc}(\rho_1, \rho_2) \stackrel{\text{def}}{=} E\left[a(k) b(k + \rho_1) c(k + \rho_2)\right], \tag{5}$$

where "$E[\cdot]$" is the expectation operator.

The bispectrum can also be defined from another perspective in terms of the signals' DFT. Let $A(\Omega), B(\Omega)$ and $C(\Omega)$ denote the DFTs of $a(k), b(k)$ and $c(k)$, receptively. The bispectrum $B_{abc}(\Omega_1, \Omega_2)$ is defined as:

$$B_{abc}(\Omega_1, \Omega_2) \stackrel{\text{def}}{=} E\left[A(\Omega_1) B(\Omega_2) C^*(\Omega_1 + \Omega_2)\right]. \tag{6}$$

It can be derived that the definitions in (4) and (6) are equivalent [39].

By definition, the bispectrum is a function of two bi-frequency variables Ω_1 and Ω_2, and it analyzes the frequency interactions between the frequency components at Ω_1, Ω_2 and $\Omega_1 + \Omega_2$ where one frequency equals to the sum of the other two. In [35], [39], [40], the properties of bispectrum (and other HOS) have been discussed in great detail. Here, we simply present two properties which will be useful for the analysis in the following paper.

1) Property 1: If the probability density functions (PDFs) of the zero-mean random processes $a(k), b(k)$ and $c(k)$ are all symmetrically distributed, then the third-order cumulant $\mathcal{R}_{abc}(\rho_1, \rho_2)$ equals to zero. According to (4), the bispectrum $B_{abc}(\Omega_1, \Omega_2)$ also equals to zero.

The zero-mean Gaussian process is a typical kind of process with symmetric PDF, then the bispectrum of zero-mean

[1]In some literature, the authors call the definition in (4) as "cross-bispectrum," and the term "bispectrum" is used only when $a(k), b(k), c(k)$ are identical to each other. In this paper, we view the "cross-bispectrum" as the generalized definition of "bispectrum," and for the sake of simplicity, we generally call $B_{abc}(\Omega_1, \Omega_2)$ defined in (4) as "bispectrum" unless mentioned.

Gaussian processes is always zero. For speech signals, according to [41], most speech signals have asymmetric possibility density functions, so it is reasonable to analyze their bispectra. As the PDFs of the non-Gaussian interferences are unknown, their bispectra are possibly zero.

2) Property 2: The cumulant of the mixture of two statistically independent random processes equals to the sum of the cumulants of the individual random processes. According to (4), this property also holds for the bispectrum of zero-mean random processes.

This property is similar to that of power spectrum for statistically independent random processes. With this property, the analysis of mixed signal in the bispectrum domain can be simplified into analyzing the bispectra of component signals separately.

B. BPD Between a Microphone Pair

In this subsection, we analyze a pair of microphone signals in the bispectrum domain. Recall the signals received by first and mth microphone. Following (2), they are rewritten here as follows:

$$\begin{aligned} y_1(k) &= \alpha_1 s_1(k) + v_1(k) + n_1(k) \\ y_m(k) &= \alpha_m s_1(k-\tau_{m1}) + v_m(k) + n_m(k). \end{aligned} \tag{7}$$

Since $n_1(k)$ and $n_m(k)$ are zero-mean Gaussian, according to "Property 1," their bispectra are identical to zero. However, the bispectra of interference signals $v_1(k)$ and $v_m(k)$ may be non-zero. Under the assumption that the interferences are uncorrelated with speech source, according to "Property 2," the bispectrum of $y_1(k)$ and the cross-bispectrum between $y_1(k), y_m(k)$ can be derived as:

$$\begin{aligned} B_{y^{111}}(\Omega_1,\Omega_2) &= \alpha_1^3 B_{s^{111}}(\Omega_1,\Omega_2) + B_{v^{111}}(\Omega_1,\Omega_2) \\ B_{y^{1m1}}(\Omega_1,\Omega_2) &= \alpha_1^2\alpha_m B_{s^{1m1}}(\Omega_1,\Omega_2) + B_{v^{1m1}}(\Omega_1,\Omega_2), \end{aligned} \tag{8}$$

where $B_{x^{abc}}(\Omega_1,\Omega_2)$ stands for the bispectrum of signal $x_a(k)$,$x_b(k)$ and $x_c(k)$.

As $s_m(k) = s_1(k - f_m(\varphi))$, according to (5), similar to the derivation in [42], we have:

$$\begin{aligned} \mathcal{R}_{s^{1m1}}(\rho_1,\rho_2) &= E\left[s_1(k)s_m(k+\rho_1)s_1(k+\rho_2)\right] \\ &= E\left[s_1(k)s_1\left(k+\rho_1-f_m(\varphi)\right)s_1(k+\rho_2)\right] \\ &= \mathcal{R}_{s^{111}}\left(\rho_1 - f_m(\varphi), \rho_2\right), \end{aligned} \tag{9}$$

where $\mathcal{R}_{s^{abc}}(\rho_1,\rho_2)$ stands for the cumulant of zero-mean signal $s_a(k)$,$s_b(k)$ and $s_c(k)$. According to the property of 2-D DFT and the definition of bispectrum in (4), the following relationship holds for $B_{s^{1m1}}(\Omega_1,\Omega_2)$ and $B_{s^{111}}(\Omega_1,\Omega_2)$:

$$B_{s^{1m1}}(\Omega_1,\Omega_2) = B_{s^{111}}(\Omega_1,\Omega_2)e^{j\Omega_1 f_m(\varphi)}. \tag{10}$$

In each bispectrum unit (Ω_1,Ω_2), we define the BPD between $B_{y^{111}}(\Omega_1,\Omega_2)$ and $B_{y^{1m1}}(\Omega_1,\Omega_2)$ as follows:

$$\Upsilon_{m1}(\Omega_1,\Omega_2) \stackrel{\text{def}}{=} \frac{B_{y^{1m1}}(\Omega_1,\Omega_2)B^*_{y^{111}}(\Omega_1,\Omega_2)}{|B_{y^{1m1}}(\Omega_1,\Omega_2)||B_{y^{111}}(\Omega_1,\Omega_2)|} \tag{11}$$

Substituting (8) and (10) into (11), the BPD can be expressed as:

$$\Upsilon_{m1}(\Omega_1,\Omega_2) = e^{j\Omega_1 f_m(\varphi)}\kappa_m(\Omega_1,\Omega_2) \tag{12}$$

where

$$\kappa_m(\Omega_1,\Omega_2) = \frac{\alpha_1^5\alpha_m|B_{s^{111}}|^2 + B_{v^{111}}B^*_{v^{1m1}}e^{-j\Omega_1 f_m(\varphi)}}{|\alpha_1^5\alpha_m|B_{s^{111}}|^2 e^{j\Omega_1 f_m(\varphi)} + B_{v^{111}}B^*_{v^{1m1}}|}, \tag{13}$$

In the right side of (13), "(Ω_1,Ω_2)" is omitted for simplicity.

The BPD in (12) is a complex number, which is expressed as a product of two complex terms. $f_m(\varphi)$ in the first term is determined by the DOA of speech source, sampling rate and array geometry. Then, for fixed sampling rate and array geometry, if ignoring the effect of attenuation factors, $e^{j\Omega_1 f_m(\varphi)}$ is only related to the DOA of the speech source and the bi-frequency Ω_1, and we call it the "speech DOA cue." The second term $\kappa_m(\Omega_1,\Omega_2)$ is related to the bispectra of the non-Gaussian interference signals. It indicates how much the speech DOA cue is affected by non-Gaussian interferences. Obviously, in pure speech units, $\kappa_m(\Omega_1,\Omega_2) = 1$, the BPD equals to the speech DOA cue.

From (8) and (10), the speech bispectrum components in $B_{y^{111}}(\Omega_1,\Omega_2)$ and $B_{y^{1m1}}(\Omega_1,\Omega_2)$ differ only up to a scale factor and phase shift. Therefore, these speech components have the same distributions in the bispectrum amplitude. As the interferences are also directional, even though the relationships between interference components are not derived here, we similarly deduce that each interference component has the same distribution in $B_{y^{111}}(\Omega_1,\Omega_2)$ and $B_{y^{1m1}}(\Omega_1,\Omega_2)$. As a result, if a bispectrum unit (Ω_1,Ω_2) of $B_{y^{111}}(\Omega_1,\Omega_2)$ is speech-dominated, the same unit in $B_{y^{1m1}}(\Omega_1,\Omega_2)$ must be also speech-dominated, and vice versa.

In speech-dominated bispectrum units, we assume that the interference bispectra $B_{v^{111}}(\Omega_1,\Omega_2)$ and $B_{v^{1m1}}(\Omega_1,\Omega_2)$, which are complex numbers, become close to 0. Then, both the numerator and denominator of $\kappa_m(\Omega_1,\Omega_2)$ in (13) approximate $\alpha_1^5\alpha_m|B_{s^{111}}|^2$, and $\kappa_m(\Omega_1,\Omega_2)$ becomes a complex number close to 1. Then, in such units, the BPD approximates the real speech DOA cue.

One interesting property which can be seen from (12) is that, although the BPD is defined in each bispectrum unit (Ω_1,Ω_2), actually, the speech DOA cue in BPD is not a function of the bi-frequency Ω_2. In other words, as long as two bispectrum units have the same Ω_1, whatever the values of their Ω_2 are, they have redundant speech DOA cues in the BPD. The redundancy of speech DOA cues in the BPD is illustrated in Fig. 2.

For a pair of signals in which one is a delayed version of the other, it is more conventional to compute the phase difference in the frequency domain. However, if the signals are polluted by noises, the bispectrum phase difference has several advantages over the frequency phase difference. One advantage is that, in theory, the effect of zero-mean Gaussian noise has been removed in the bispectrum domain. Another advantage is that the redundancy of speech DOA cue in the BPD helps to improve the robustness against interferences. As the bispectra of the speech and interferences distribute differently, in some speech units, the speech DOA cues may be severely polluted, nevertheless,

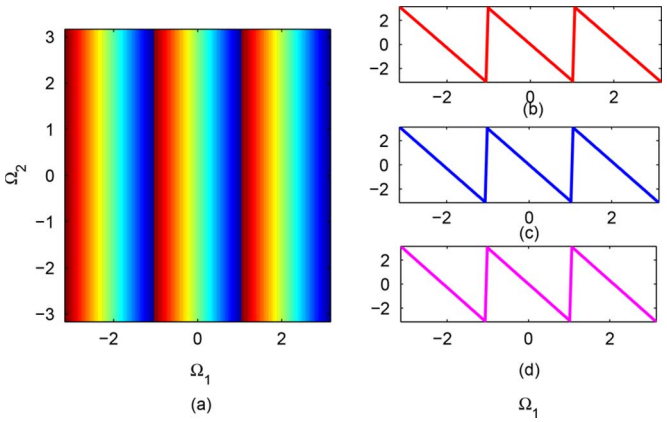


Fig. 2. Illustration of the redundancy of speech DOA cues in the BPD between two clean signals with one as the delayed version of the other. (a) The angle spectrum of speech DOA cues in the BPD. The angle of the speech DOA cues in the BPD for different Ω_2's: (b) $\Omega_2 = \frac{3\pi}{4}$, (c) $\Omega_2 = \frac{\pi}{2}$, (d) $\Omega_2 = 0$.

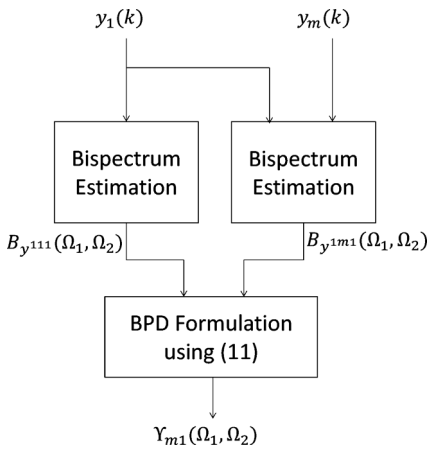


Fig. 3. The procedure of computing BPD $\Upsilon_{m1}(\Omega_1,\Omega_2)$ using measured data of $y_1(k)$ and $y_m(k)$.

they can be re-found in other speech-dominated units with the same Ω_1. In contrast, for the frequency domain representation, once the cue in one speech band is polluted by the interference, it cannot be found elsewhere. Therefore, this property of BPD helps to improve the performance of the algorithm in interference-existing-scenarios. Furthermore, as the speech DOA cues are repeatedly expressed, we can check whether the candidate DOA matches the real one by utilizing more than only one set of DOA cues, thus it can be viewed that the redundancy brings more observations for DOA estimation.

C. Numerical Computation of BPD From Measured Data

One problem in practice is how to estimate the BPD of a microphone pair given a finite set of measurements. Following the definition of BPD in (11), the procedure of computing BPD between $y_1(k), y_m(k)$ using the measured data is shown in the Fig. 3. In the procedure, the key issue is the "bispectrum estimation," in which the bispectra $B_{y^{111}}(\Omega_1,\Omega_2)$ and $B_{y^{1m1}}(\Omega_1,\Omega_2)$ are estimated using finite observations of $y_1(k), y_m(k)$. It should be noted that $B_{y^{111}}(\Omega_1,\Omega_2)$ is a special case of $B_{y^{1m1}}(\Omega_1,\Omega_2)$ when $m = 1$. Therefore, only the method of computing $B_{y^{1m1}}(\Omega_1,\Omega_2)$ is presented here.

Conventional bispectrum estimation approaches include "direct" and "indirect" methods [39], which can be seen as approximations of the two different definitions of bispectrum (6) and (4) in Section III-A. As the DFT of the signal can be efficiently estimated by the Fast Fourier Transform (FFT), we adopt the direct method for bispectrum estimation. The implementation details of the direct method can be found in [39]. In this method, in order to reduce the variance of bispectrum estimation, the "segment-average" strategy is commonly used. In this strategy, the given set of measurements are firstly segmented into small overlapping segments, then the estimated bispectra in these overlapping segments are averaged to obtain a final bispectrum estimation.

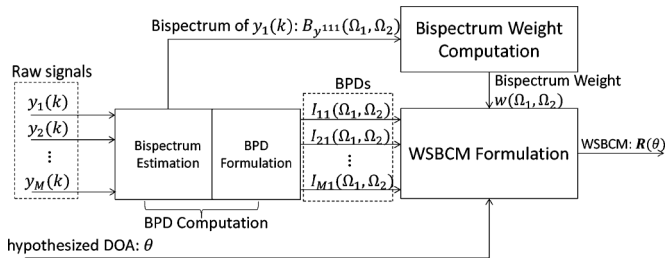


Fig. 4. The flowchart of computing the WBSCM.

In our problem, the DOA estimation is expected to be conducted in the frame level, then the data length of one frame may be not long enough for segmentation. Only a few segments or even one segment can be segmented from one frame of data. In such case, in addition to the "segment-average" strategy, we also smooth over the bispectrum estimation results of consecutive frames to reduce the estimation variance. Assuming we have obtained the bispectrum estimation $\tilde{B}^{(t)}_{y^{1m1}}(\Omega_1,\Omega_2)$ of $B_{y^{1m1}}(\Omega_1,\Omega_2)$ in frame t using the direct method. Then the final bispectrum estimation is updated as:

$$\hat{B}^{(t)}_{y^{1m1}}(\Omega_1,\Omega_2) = \beta_1 \hat{B}^{(t-1)}_{y^{1m1}}(\Omega_1,\Omega_2) + (1-\beta_1)\tilde{B}^{(t)}_{y^{1m1}}(\Omega_1,\Omega_2), \tag{14}$$

where the superscript $(\cdot)^{(t)}$ indicates the time frame t, and $\hat{(\cdot)}$ stands for the estimated value. $\beta_1 \in [0,1)$ is a smoothing factor. Although the larger value of β_1 helps to reduce the estimation variance, smaller value makes the estimation can more effectively track the update of bispectrum. In this paper, as a compromise, the β_1 is empirically set to be 0.7. With the bispectrum estimation $\hat{B}^{(t)}_{y^{1m1}}(\Omega_1,\Omega_2)$ and $\hat{B}^{(t)}_{y^{111}}(\Omega_1,\Omega_2)$, the BPD can be computed according to (11).

IV. Weighted Bispectrum Spatial Correlation Matrix

When multiple microphones are available, the BPDs between multiple microphones and the reference one can be computed, and it is possible to utilize multiple BPDs to improve the DOA estimation performance. The problem is how to utilize these BPDs in a proper way. In this section, we integrate the BPDs of multiple microphones into a compact mathematical expression called "WBSCM," which will be used by the DOA estimator in the next section. The flowchart of computing WBSCM using the raw data of microphone array is shown in Fig. 4. It includes three main parts. The first part involves computing the BPDs between multiple microphones and the reference one, and it has been introduced in Section III-C. The second part involves computing the bispectrum weights using the bispectrum of the first

microphone. In the third part, the computed BPDs and bispectrum weights are used to formulate the WBSCM for a hypothesized DOA. In the following, we will begin with the method for bispectrum weights computation, then the formulation of the WBSCM will be introduced.

A. Bispectrum Weights

According to (11), the real DOA cue is close to BPD only in speech-dominated bispectrum units, therefore, it is better to pick out only these units for DOA estimation. One common way to achieve this is to utilize a set of non-negative bispectrum weights, and these weights are expected to have large values in speech bispectrum units, and zero values in non-speech ones.

Actually, the bispectrum weight, which can be viewed as an indicator of how much the speech signal is polluted in one bispectrum unit, is analogous to the concept of SNR in the frequency domain. Therefore, the ideas for SNR estimation can be exploited to compute the bispectrum weights. Here, a "decision-directed" method for computing bispectrum weights is proposed, which is similar to the decision-directed *a priori* SNR estimator proposed by Ephraim and Malah [43].

As is analyzed in Section III-B, both the speech and interference have the same bispectrum amplitude distributions in $B_{y^{1m1}}(\Omega_1,\Omega_2)$ and $B_{y^{111}}(\Omega_1,\Omega_2)$, therefore, it is the same to use either $B_{y^{111}}(\Omega_1,\Omega_2)$ or $B_{y^{1m1}}(\Omega_1,\Omega_2)$ for bispectrum weight calculation. Here, the $B_{y^{111}}(\Omega_1,\Omega_2)$ is chosen. Similar to the definition of SNR, we define the local *a priori* bispectrum signal-to-interference ratio (BSIR) $\xi(\Omega_1,\Omega_2)$ and *a posteriori* BSIR $\gamma(\Omega_1,\Omega_2)$ of the bispectrum unit (Ω_1,Ω_2) as:

$$\xi(\Omega_1,\Omega_2) \stackrel{\text{def}}{=} \frac{|B_{s^{111}}(\Omega_1,\Omega_2)|^2}{\lambda_v(\Omega_1,\Omega_2)}, \tag{15}$$

$$\gamma(\Omega_1,\Omega_2) \stackrel{\text{def}}{=} \frac{|B_{y^{111}}(\Omega_1,\Omega_2)|^2}{\lambda_v(\Omega_1,\Omega_2)}, \tag{16}$$

where $\lambda_v(\Omega_1,\Omega_2) \stackrel{\text{def}}{=} E\{|B_{v^{111}}(\Omega_1,\Omega_2)|^2\}$ is the estimation of interference bispectrum power, which is initialized and updated during the period of silence. We assume that the speech is not active in the initial short-time period (assumed to be 0.3s long in this paper), then $\lambda_v(\Omega_1,\Omega_2)$ can be initialized as the average value of the bispectrum power computed within the initial frames:

$$\lambda_v(\Omega_1,\Omega_2) = \frac{1}{K}\sum_{i=1}^{K}\left|\hat{B}^{(i)}_{y^{111}}(\Omega_1,\Omega_2)\right|^2, \tag{17}$$

where K is the number of initial frames. The update of $\lambda_v(\Omega_1,\Omega_2)$ will be introduced later in this section in (26).

Assuming the speech and interferences are uncorrelated, it is clear that:

$$\xi(\Omega_1,\Omega_2) = \gamma(\Omega_1,\Omega_2) - 1. \tag{18}$$

Following (15) and (18), the *a priori* BSIR $\xi(\Omega_1,\Omega_2)$ is estimated as:

$$\begin{aligned}\hat{\xi}^{(t)}(\Omega_1,\Omega_2) =& \beta_2 \frac{|\hat{B}^{(t-1)}_{s^{111}}(\Omega_1,\Omega_2)|^2}{\lambda_v^{(t)}(\Omega_1,\Omega_2)} \\ &+ (1-\beta_2)P\left[\gamma^{(t)}(\Omega_1,\Omega_2) - 1\right],\end{aligned} \tag{19}$$

where $P[\cdot]$ denotes half-wave rectification operator, which ensures the positiveness of the estimated *a priori* BSIR, and it is defined by:

$$P[x] = \begin{cases} x & \text{if x} \geq 0 \\ 0 & \text{otherwise} \end{cases}. \tag{20}$$

The β_2 is a smoothing factor, which is similar with that used in the *a priori* SNR estimation [43]. For *a priori* SNR estimation, typical values of the smoothing factor are in the range of [0.92, 0.98], and larger values often lead to more noise reduction and speech distortion [44], [45]. Here, as the speech distortion is beyond our concern, and the aim is to find reliable bispectrum units for DOA estimation, the smoothing factor is set to be 0.98.

The *a priori* BSIR estimator in (19) is a "decision-directed" estimator, since the $\hat{\xi}^{(t)}(\Omega_1,\Omega_2)$ is updated based on the previous estimation of the speech bispectrum $\hat{B}^{(t-1)}_{s^{111}}(\Omega_1,\Omega_2)$. The $|\hat{B}^{(t-1)}_{s^{111}}(\Omega_1,\Omega_2)|^2$ in the right side of (19) is unknown, it can be approximated as:

$$\begin{aligned}\left|\hat{B}^{(t-1)}_{s^{111}}\right|^2 &= \frac{\left|\hat{B}^{(t-1)}_{s^{111}}\right|^2}{\left|\hat{B}^{(t-1)}_{y^{111}}\right|^2} * \left|\hat{B}^{(t-1)}_{y^{111}}\right|^2 \\ &= \frac{\left|\hat{B}^{(t-1)}_{s^{111}}\right|^2}{\left|\hat{B}^{(t-1)}_{s^{111}}\right|^2 + \lambda_v^{(t)}} * \left|\hat{B}^{(t-1)}_{y^{111}}\right|^2 \\ &= \frac{\hat{\xi}^{(t-1)}}{\hat{\xi}^{(t-1)}+1} * \left|\hat{B}^{(t-1)}_{y^{111}}\right|^2.\end{aligned} \tag{21}$$

Substituting (21) into (19), we have:

$$\begin{aligned}\hat{\xi}^{(t)} &= \beta_2 \frac{\hat{\xi}^{(t-1)}}{\hat{\xi}^{(t-1)}+1} * \frac{\left|\hat{B}^{(t-1)}_{y^{111}}\right|^2}{\lambda_v^{(t)}} + (1-\beta_2)P\left[\hat{\gamma}^{(t)} - 1\right] \\ &= \beta_2 \hat{\gamma}^{(t-1)} \frac{\hat{\xi}^{(t-1)}}{\hat{\xi}^{(t-1)}+1} + (1-\beta_2)P\left[\hat{\gamma}^{(t)} - 1\right].\end{aligned} \tag{22}$$

In (21) and (22), the symbol "(Ω_1,Ω_2)" is omitted for simplicity.

We should point out that, although the estimated *a priori* BSIR $\hat{\xi}^{(t)}(\Omega_1,\Omega_2)$ reflects how much the speech signal is polluted by the interference in one bispectrum unit, not all units with high *a priori* BSIRs are the target units to be selected. In a unit where both the speech and interference distribute little or are absent, even though the bispectrum powers of the speech and interference are both small, the relative value between them may be large, thus the *a priori* BSIR may be also high. As we aim to select "speech-dominated" bispectrum units, it is better to find out the speech units in the first step, and then pick out the speech-dominated ones.

Following (21), with the estimated BSIR $\hat{\xi}^{(t)}(\Omega_1,\Omega_2)$, it is straightforward to estimate the speech bispectrum power in current frame, i.e.:

$$\left|\hat{B}^{(t)}_{s^{111}}(\Omega_1,\Omega_2)\right|^2 = \frac{\hat{\xi}^{(t)}(\Omega_1,\Omega_2)}{\hat{\xi}^{(t)}(\Omega_1,\Omega_2)+1} * \left|\hat{B}^{(t)}_{y^{111}}(\Omega_1,\Omega_2)\right|^2. \tag{23}$$

The estimations of the speech bispectrum power in different units provide a description of the speech bispectrum distribu-

tion. As $|\hat{B}^{(t)}_{s^{111}}(\Omega_1,\Omega_2)|^2$ will be large in speech units, and small in non-speech ones, we can use $|\hat{B}^{(t)}_{s^{111}}(\Omega_1,\Omega_2)|^2$ to indicate the presence of speech in a unit, and select the speech units as the ones with large $|\hat{B}^{(t)}_{s^{111}}(\Omega_1,\Omega_2)|^2$.

Then speech-dominated units are further picked out from these selected speech units. According to the definition of $\xi(\Omega_1,\Omega_2)$ in (15), it can be seen that if a unit is dominated by speech, $\xi(\Omega_1,\Omega_2)$ will be larger than 1. Here, a threshold $\zeta \geq 1$ is applied to $\xi(\Omega_1,\Omega_2)$ for selecting speech-dominated units. On one hand, we intend to select more speech-dominated units for DOA estimation, while on the other hand, these selected units are expected to be as clean as possible. Obviously, the larger ζ is, the less polluted bispectrum units are selected, but the amount of selected units gets fewer. Therefore, there is a trade-off with respect to the choice of ζ. In high SNR conditions, as the BSIRs in speech-dominated units are also high, a higher value of ζ can be adopted to improve the performance, however, when the noise level increases, ζ should be set close to 1. In our experiment, considering different noisy conditions, we empirically choose the value of ζ to be 1.5. Selecting the speech-dominated units can be expressed as follows:

$$G(\Omega_1,\Omega_2) = \begin{cases} \xi(\Omega_1,\Omega_2) - \zeta & \text{if } \xi(\Omega_1,\Omega_2) \geq \zeta \\ 0 & \text{otherwise.} \end{cases} \tag{24}$$

We define the bispectrum weight as the combination of the above two steps for selecting speech-dominated bispectrum units:

$$w(\Omega_1,\Omega_2) \stackrel{\text{def}}{=} \left|\hat{B}^{(t)}_{s^{111}}(\Omega_1,\Omega_2)\right|^2 G(\Omega_1,\Omega_2), \tag{25}$$

Obviously the bispectrum weights will be equal to zero in non-speech-dominated units.

Fig. 5 shows an example of the calculated bispectrum weights (plus a small positive number and transform to the log-scale), the log-scale estimated *a priori* BSIR and the log-scale amplitude of the bispectra corresponding to the noisy speech, clean speech and interference in one frame, respectively. It can be observed that although the speech bispectrum units are less visible in the noisy speech bispectrum, these units are effectively recovered in the bispectrum weights, and the effect of the interference has been almost totally removed.

The $G(\Omega_1,\Omega_2)$ in (24) can also be used for a speech activity detection (VAD) in the bispectrum domain. We assume that the current frame is a noise one if $G(\Omega_1,\Omega_2) = 0$ for all bispectrum units. Then the $\lambda_v(\Omega_1,\Omega_2)$ can be updated as:

$$\lambda_v^{(t)}(\Omega_1,\Omega_2) = \beta_3 \lambda_v^{(t-1)}(\Omega_1,\Omega_2) + (1-\beta_3)\left|\hat{B}^{(t)}_{y^{111}}(\Omega_1,\Omega_2)\right|^2, \tag{26}$$

where β_3 is a smoothing factor. We assume that in the bispectrum domain, the noise signal is more stationary than the speech. In this paper, we use a low update rate of the estimated interference bispectrum power, and β_3 is empirically chosen to be 0.95.

B. Formulation of WBSCM

In each bispectrum unit (Ω_1,Ω_2), the BPDs between multiple microphones and the reference one can be computed. Actually, we can see that the expressions of speech DOA cues for different microphones, which are defined in (13), are much similar to the expressions of complex signals received by different microphones in the classical narrow-band MUSIC algorithm. Therefore, one may solve the DOA estimation problem by narrowband methods such as MUSIC in each (Ω_1,Ω_2), and finally combine these estimations using the computed bispectrum weights. However, it is obviously time-consuming. In this section, we integrate these BPDs into a compact matrix called "WBSCM," which reflects the spatial correlations between the phase aligned BPDs for a hypothesized DOA. It enables bispectrum weighting to highlight the effects of speech-dominated bispectrum units so as to further improve the robustness against the noise, and shows an interesting property only when the hypothesized DOA equals to the true one.

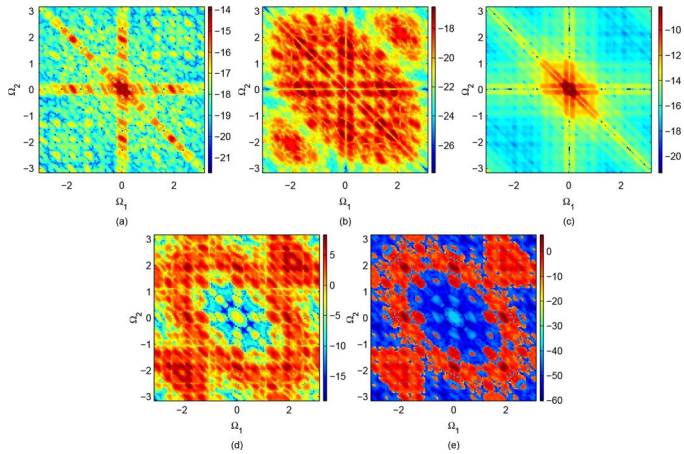


Fig. 5. Example of the calculated bispectrum weights in one frame. The noise environment: car interior noise, SIR $=$ 5 dB and $\mathrm{RT}_{60} = 250$ ms. The log-scale bispectrum amplitude of (a) the noisy speech signal, (b) the clean speech signal, and (c) the interference signal. The log-scale estimated *a priori* BSIR is shown in (d), and the log-scale bispectrum weights is shown in (e).

In the bispectrum unit (Ω_1,Ω_2), we define a BPD vector which consists of the BPDs between all microphones and the reference one as:

$$\boldsymbol{\Upsilon}(\Omega_1,\Omega_2) \stackrel{\text{def}}{=} [\Upsilon_{11}(\Omega_1,\Omega_2), \Upsilon_{21}(\Omega_1,\Omega_2), \ldots, \Upsilon_{M1}(\Omega_1,\Omega_2)]^T. \tag{27}$$

The $\boldsymbol{\Upsilon}(\Omega_1,\Omega_2)$ is a complex-value vector. Obviously, an explicit theoretical expression of the BPD vector can be derived according to (12):

$$\boldsymbol{\Upsilon}(\Omega_1,\Omega_2) = \Big[\kappa_1(\Omega_1,\Omega_2), \kappa_2(\Omega_1,\Omega_2)e^{j\Omega_1 f_2(\varphi)}, \ldots, \kappa_M(\Omega_1,\Omega_2)e^{j\Omega_1 f_M(\varphi)}\Big]^T \tag{28}$$

It has been discussed in Section III-B that in pure speech units, κ_m for $m = 1,\ldots,M$ equals to 1, and the BPD equals to the real DOA cue. In practice, due to the noise and bispectrum estimation error, this ideal case rarely happens, therefore, we express the BPD vector as the sum of the ideal one and an noise vector:

$$\boldsymbol{\Upsilon}(\Omega_1,\Omega_2) = \left[1, e^{j\Omega_1 f_2(\varphi)}, \ldots, e^{j\Omega_1 f_M(\varphi)}\right]^T + \boldsymbol{\varepsilon}(\Omega_1,\Omega_2), \tag{29}$$

where $\boldsymbol{\varepsilon}(\Omega_1,\Omega_2)$ is the noise vector with the mth element written as: $\boldsymbol{\varepsilon}_m(\Omega_1,\Omega_2) = (\kappa_m(\Omega_1,\Omega_2) - 1)e^{j\Omega_1 f_m(\varphi)}$.

Suppose the true DOA of speech source is known exactly, then we can use a "BPD compensation vector" to totally compensate the speech DOA cues in the BPD vector. Inspired by the expression of speech DOA cue in (11), the BPD compensation vector for a hypothesized DOA θ is defined as:

$$\boldsymbol{C}(\theta,\Omega_1) \stackrel{\text{def}}{=} \left[1, e^{-j\Omega_1 f_2(\theta)}, \ldots, e^{-j\Omega_1 f_M(\theta)}\right]^T. \tag{30}$$

Then we compensate the BPDs using the defined $\boldsymbol{C}(\theta,\Omega_1)$ as follows:

$$\boldsymbol{\Upsilon}^c(\theta,\Omega_1,\Omega_2) \stackrel{\text{def}}{=} \boldsymbol{\Upsilon}(\Omega_1,\Omega_2) \odot \boldsymbol{C}(\theta,\Omega_1), \tag{31}$$

where $\boldsymbol{\Upsilon}^c(\theta,\Omega_1,\Omega_2)$ is the $M \times 1$ phase-compensated BPD vector, and the symbol "$\odot$" stands for the element-wise multiplication operator. Substituting (29) and (30) into (31), we have:

$$\begin{aligned}\boldsymbol{\Upsilon}^c(\theta,\Omega_1,\Omega_2) &= \left[1, e^{j\Omega_1(f_2(\varphi)-f_2(\theta))}, \ldots, e^{j\Omega_1(f_M(\varphi)-f_M(\theta))}\right]^T \\ &\quad + \boldsymbol{\varepsilon}(\Omega_1,\Omega_2) \odot \boldsymbol{C}(\theta,\Omega_1) \\ &= \boldsymbol{\Gamma}(\theta,\Omega_1) + \boldsymbol{\varepsilon}^c(\theta,\Omega_1,\Omega_2),\end{aligned} \tag{32}$$

where $\boldsymbol{\varepsilon}^c(\theta,\Omega_1,\Omega_2)$ is the phase-compensated noise vector.

With the bispectrum weights computed using (25), we define the WBSCM $\boldsymbol{R}(\theta)$ which contains the spatial correlation information of the phase aligned BPDs for θ as follows:

$$\boldsymbol{R}(\theta) \stackrel{\text{def}}{=} E\left\{\sum_{(\Omega_1,\Omega_2)} w(\Omega_1,\Omega_2)\left[\boldsymbol{\Upsilon}^c(\theta,\Omega_1,\Omega_2)\right]\left[\boldsymbol{\Upsilon}^c(\theta,\Omega_1,\Omega_2)\right]^H\right\}. \tag{33}$$

The $\boldsymbol{R}(\theta)$ is only a function of the hypothesized DOA θ, and it can be viewed as a weighted sum of the correlation matrices of the multiple phase-aligned BPDs in different bispectrum units. As the $\boldsymbol{\Upsilon}^c(\theta,\Omega_1,\Omega_2)$ is a complex-valued $M \times 1$ vector, the $\boldsymbol{R}(\theta)$ is an $M \times M$ hermitian matrix.

We assume that in speech-dominated bispectrum units, the phase-compensated noise vector noise vector $\boldsymbol{\varepsilon^c}(\theta,\Omega_1,\Omega_2)$ can be ignored. Furthermore, the bispectrum weight $w(\Omega_1,\Omega_2)$ is expected be zero in non-speech-dominated units, then substituting (32) into (33), the WBSCM can be rewritten as:

$$\boldsymbol{R}(\theta) = E\left\{\sum_{(\Omega_1,\Omega_2)\in\boldsymbol{\Omega_S}} \left[w(\Omega_1,\Omega_2)\boldsymbol{\Gamma}(\Omega_1,\Omega_2)\boldsymbol{\Gamma}(\Omega_1,\Omega_2)^H\right]\right\}, \tag{34}$$

where $\boldsymbol{\Omega_S}$ denotes the set of speech-dominated units in the bispectrum which are selected by the bispectrum weights. Once θ matches φ, according to (32), $\boldsymbol{\Gamma}(\theta,\Omega_1)$ becomes an real vector with all elements equal to 1, i.e., $\boldsymbol{\Gamma}(\theta,\Omega_1) = \boldsymbol{e} = [1,1,\ldots,1]^H$, indicating the speech DOA cues are phased aligned in BPDs. In this case $\boldsymbol{\Gamma}(\theta,\Omega_1)$ is no longer related to Ω_1, then

$$\begin{aligned}\boldsymbol{R}(\theta) &= E\left\{\sum_{(\Omega_1,\Omega_2)\in\boldsymbol{\Omega_S}} w(\Omega_1,\Omega_2)\right\}(\boldsymbol{e}^H\boldsymbol{e}) \\ &= \eta\cdot(\boldsymbol{e}^H\boldsymbol{e}) \\ &= \eta\cdot\begin{bmatrix}1 & \ldots & 1\\ \vdots & \ddots & \vdots\\ 1 & \ldots & 1\end{bmatrix},\end{aligned} \tag{35}$$

where $\eta = E\{\sum_{(\Omega_1,\Omega_2)\in\boldsymbol{\Omega_S}} w(\Omega_1,\Omega_2)\}$ is a real constant given $w(\Omega_1,\Omega_2)$ in a certain frame. Therefore, we note that, ideally, if $\theta = \varphi, \boldsymbol{R}(\theta)$ becomes a matrix with all elements equal to a real constant. Otherwise, it is just an ordinary hermitian matrix.

The numerical computation of WBSCM can be performed by following the definition in (33). In practice, let us use the symbol $\tilde{\boldsymbol{R}}^{(t)}(\theta)$ to donate the value of "$\sum_{(\Omega_1,\Omega_2)} w(\Omega_1,\Omega_2)[\boldsymbol{\Upsilon}^c(\theta,\Omega_1,\Omega_2)][\boldsymbol{\Upsilon}^c(\theta,\Omega_1,\Omega_2)]^H$" computed in the time frame t, then the expectation operator can be implemented by recursively smoothing over time:

$$\hat{\boldsymbol{R}}^{(t)}(\theta) = \beta_4\hat{\boldsymbol{R}}^{(t-1)}(\theta) + (1-\beta_4)\tilde{\boldsymbol{R}}^{(t)}(\theta), \tag{36}$$

where $\hat{\boldsymbol{R}}^{(t)}(\theta)$ is the WBSCM finally computed in the t-th frame, and $\beta_4 \in [0,1)$ is a smoothing factor. The β_4 plays a role similar with the β_1 in (14). Since in general, the speech signal is much more stationary in spatial domain than in time domain, β_4 can be chosen much more close to 1 computed with β_1. A larger value of β_4 helps to reduce the estimation fluctuation of WBSCM, however, too large β_4 reduces the update speed of WBSCM. We can not avoid the cases when WBSCM contains more bispectrum spatial correlation information of the interference than the speech, which lead to wrong estimations, especially in highly noisy conditions. Then too large β_4 makes the algorithm can not refind the speech source rapidly in such cases. In this paper, as a compromise between the performances in different conditions, the β_4 is empirically set to be 0.98.

V. DOA Estimator

Although in theory, the WBSCM will be a matrix with all elements equal to a real constant if $\theta = \varphi$, in reality, as the noise vector will not be exactly zero, this ideal case rarely happens. Therefore, we need to find measurements to indicate how much the WBSCM approximates the ideal case for a given θ.

In this section, we propose a DOA estimator based on the eigenvalue analysis of the WBSCM. It is clear that if all the elements in the WBSCM are equal, the rank of WBSCM will be 1. Otherwise, the WBSCM is an ordinary hermitian matrix, and it will be full-rank. Let's perform eigenvalue decomposition of $\boldsymbol{R}(\theta)$ and let $\lambda_1(\theta),\ldots,\lambda_M(\theta)$ denote the eigenvalues of $\boldsymbol{R}(\theta)$, which are complex numbers, with their absolute values sorted in a decreasing order, i.e., $|\lambda_1(\theta)| \geq |\lambda_2(\theta)| \geq \ldots \geq |\lambda_M(\theta)|$. Obviously, if $\boldsymbol{R}(\theta)$ is of rank 1, $|\lambda_2(\theta)| = \ldots = |\lambda_M(\theta)| = 0$. Therefore, if we form the following cost function

$$\boldsymbol{J}(\theta) \stackrel{\text{def}}{=} \frac{1}{\sum_{i=2}^{M}|\lambda_i(\theta)|}, \tag{37}$$

the cost function reaches the maximum if $\theta = \hat{\theta}$. Then the estimated DOA $\tilde{\theta}$ is calculated as:

$$\tilde{\theta} = arg\max_{\theta} \boldsymbol{J}(\theta). \tag{38}$$

VI. Experiment

In this section, in order to evaluate the performance of the proposed algorithm and other comparison algorithms, we conduct experiments in both simulated and real room environments.

A. Experimental Setup

1) Simulated Room Environment: In the simulated room environment, the sound sources and the microphone array are assumed to be located in a rectangular reverberant room with dimensions: length $=$ 600 cm, width $=$ 400 cm, and height $=$ 300 cm ($x \times y \times z$). We use the image-source method [46] to simulate the reverberant environment, and a Matlab code implementation[2] of this method is utilized for generating the room impulse responses (RIRs) from sound sources to microphones.

We employ a ULA consisting of two to six omni-directional microphones to capture the signals from sound sources, and the spacing between two adjacent microphones is 10 cm. The microphone array is located at the geometric center of the rectangular room, with all elements in the horizontal plane, and the array orientation parallel to the long edge of the horizontal plane. For the speech and interference sources, it is set that all sound sources are situated on a horizontal plane (x, y, 150 cm) with distance 190 cm to the center of the microphone array, and the acoustic signals emitted from these sources are sampled with 8 kHz sampling rate and 16-bit resolution. In all simulated situations, the reverberation time RT_{60} of the room is set to be 250 ms.

2) Real Room Environment: The Multichannel Impulse Response Database (MIRD)[3], which contains real RIRs measured in the speech & acoustic lab at the Bar-Ilan University (BIU) [47], is exploited to generate the multichannel signals in real room conditions. The room size of the BIU speech & acoustic lab is 600 cm $\times$ 600 cm $\times$ 240 cm, and an eight-element microphone array is utilized to capture the sounds. All measurements in the database are sampled with 48 kHz sampling rate and 24-bit resolution. Different reverberation times, microphone spacings, and source-array distances are configured, therefore several subsets are included in the database. For more details of different configurations the readers can refer to [47]. In our experiment, we choose a moderately reverberant room environment with RT_{60} as 360 ms. The spacing between two adjacent microphones is 8 cm, thus a ULA is adopted. The sound sources are placed at a distance of 200 cm to the center of the microphone array, and the DOAs are from -90° to 90° with step size as 15°.

3) Speech and Interference Source Signals: A male speech of 30 seconds is used as the speech source. We utilize four different types of noises (white Gaussian noise, car interior noise, F16 cockpit noise and speech babble noise) drawn from Noisex92 [48] as interference signals. All source signals are sampled with 8 kHz. The spectrograms of the speech and interferences are shown in Figs. 6 and 7, respectively.

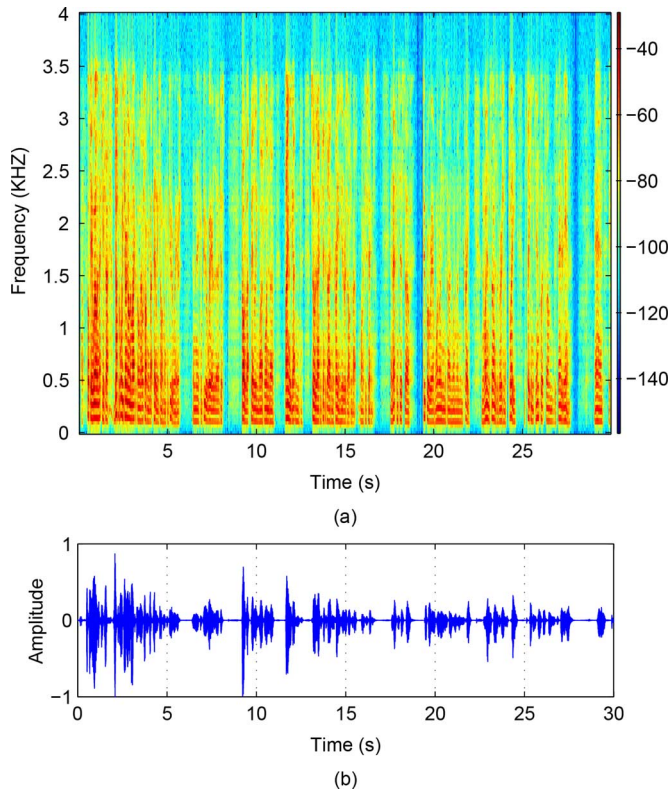


Fig. 6. (a) Spectrogram and (b) time-domain waveform of pure speech signal.

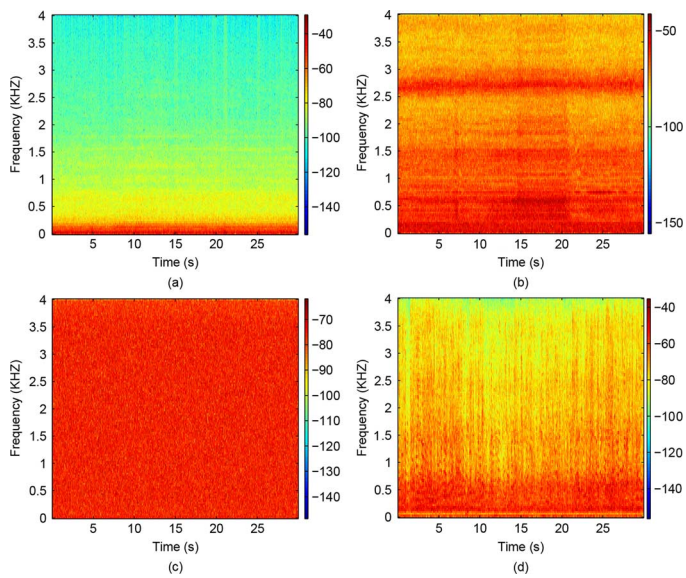


Fig. 7. Spectrograms of different interference signals: (a) car interior noise, (b) F16 cockpit noise, (c) white Gaussian noise, and (d) speech babble noise.

B. Evaluated Algorithms

Three other methods capable of multiple microphone DOA estimation are used for comparison. These methods include the well-known SRP-PHAT method which is originally introduced by J. DiBiase [12], the broadband MUSIC algorithm proposed by J. P. Dmochowski and J. Benesty, *et al.* [11], and the interference robust DOA estimation method in [26]. In the SRP-PHAT and broadband MUSIC algorithm, the cross-spectrum between two microphone signals, which is defined as

$$G_{y_i,y_j}(f) = E\left[Y_i(f)Y_j^*(f)\right], \tag{39}$$

need to be estimated, where $Y_i(f)$ denotes the DFT of the received signal of the ith microphone. Similar to (36), the estimation is performed by recursively smoothing over the previous estimation:

$$G_{y_i,y_j}^{(t)}(f) = \beta_5 G_{y_i,y_j}^{(t-1)}(f) + (1-\beta_5)Y_i^{(t)}(f)\left(Y_j^{(t)}(f)\right)^*, \tag{40}$$

where $\beta_5 = 0.95$ is a smoothing factor.

[2]The code can be found at: http://www.eric-lehmann.com/ism_code.html

[3]The MIRD can be downloaded at: http://www.ind.rwth-aachen.de/en/research/tools-downloads/multichannel-impulse-response-database/

For all evaluated algorithms, the analysis frame size is set to be 512 samples with 50% overlap. The DOA cost functions are computed on a 5° grid, then sinc-interpolation is performed on each cost function to achieve the 1° resolution.

C. Performance Criteria

All algorithms estimate the DOA in the frame-level. Two frame level metrics, denoted as accuracy and root mean square error (RMSE), are used to evaluate the performance of different algorithms. We consider the estimation result of one frame as correct if the absolute value of the estimation error does not exceed a threshold, i.e., $|\theta(t) - \varphi| \leq e_{\text{th}}$, where the threshold e_{th} is set to be 5° here. Then accuracy and RMSE are defined as:

$$\text{Accuracy} = \frac{N_0}{N} \times 100\%, \tag{41}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\theta(i) - \varphi)^2}, \tag{42}$$

where N_0 is the number of speech frames which have the correct estimation, N is the number of total speech frames. We only consider the speech frames for evaluation, and the speech frames are labeled manually in advance on the clean speech signal. It should be pointed out that these labels are never used by any of the algorithms for evaluation.

D. Performance in Simulated Room Environments

1) Spatially White Gaussian Noise Conditions: The additive spatially white Gaussian noise condition is the most widely considered condition in traditional DOA estimation algorithms. Therefore, we first test the estimation performance in this scenario. For sake of completeness, we examine different situations where the DOA of speech source ranges from $-90°$ to $90°$ with a step size of 20°. At each DOA, the RIRs from the speech source to different microphones are simulated, and the speech signals received by microphones are generated by convolving the speech signal with these RIRs. Then, the white Gaussian noise is generated at each microphone independently, and added to the received speech signal after being scaled to control the SNR. Therefore, for each SNR, we have 7 groups of simulated signals corresponding to 7 different speech DOAs.

The algorithms are first evaluated under different SNRs. The SNR changes from -10 dB to 20 dB, with the step as 5 dB. In each SNR and simulated speech DOA condition, different algorithms are utilized to estimate the speech DOAs in the frame-level. Then for each algorithm and tested SNR, we utilize all the estimation results corresponding to 7 different simulated DOA scenarios to evaluate the overall performance.

Fig. 8 depicts the comparison results under different SNRs in spatially white Gaussian noise conditions. In Fig. 8(a), these algorithms can achieve similar accuracy in high SNR situations, but if the SNR drops below 5 dB, we can clearly observe the robustness of the proposed algorithm. Even when the SNR $= -10$ dB, the accuracy of proposed method is still higher than 80%. Furthermore, from Fig. 8(b), it can be seen that in almost all SNRs considered, the proposed algorithm yields the lowest RMSE. As the proposed method is based on the HOS, in terms of the immunity of HOS against the Gaussian noise, the robustness improvement in these set of testing conditions seems explicable. In spite of this, the improvement may also

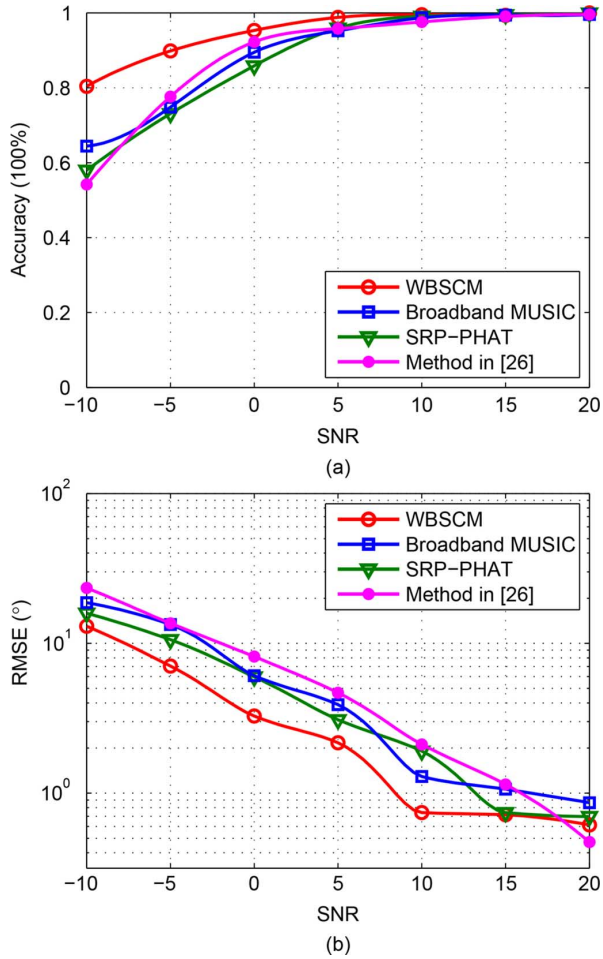


Fig. 8. (a) Estimation accuracy and (b) RMSE in spatially white Gaussian noise under different SNRs in the simulated room environment. Six microphones are used.

benefit from the redundancy provided by the BPD. Theoretically, the HOS of Gaussian noise is zero, however in practice, the bispectrum residual of Gaussian noise still exists. Although in the bispectrum domain, the effect of Gaussian noise has been much reduced, the residual still has a negative impact on the performance. In the proposed algorithm, utilizing the redundancy in BPD over one bi-frequency can be regarded as using more observed data for DOA estimation, therefore, the negative impact caused by bispectrum residual of Gaussian noise is further eliminated.

Then we investigate the performance of the evaluated algorithms as a function of the number of microphones. In this set of experiments, we fix the SNR to be 10 dB, and change the number of microphones from 2 to 6. Then for each algorithm, again, the estimation results for different DOA scenarios are combined together to achieve the overall performance evaluation. It can be seen from Fig. 9 that the performance of all algorithms generally improves as the number of microphones increases, indicating that the spatial redundancy provided by more microphones helps to improve the robustness. Nevertheless, compared with other methods, the proposed method can maintain high performance when fewer microphones are used.

2) Directional Interference Conditions: The performance of different algorithms in interference-existing scenarios is tested in this subsection. We evaluate the estimation results in different signal-to-interference ratios (SIRs) and types of interferences

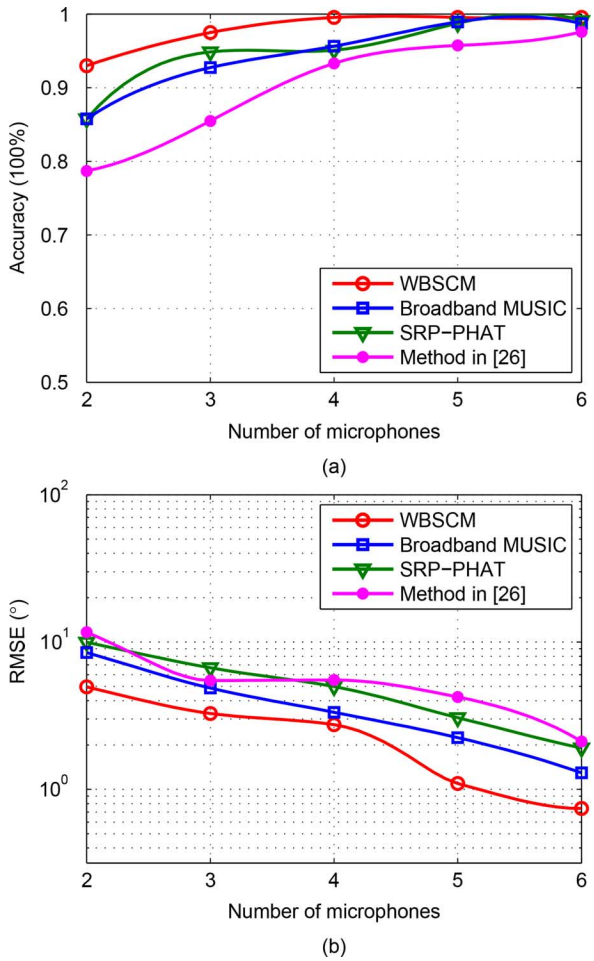


Fig. 9. (a) Estimation accuracy and (b) RMSE as a function of number of microphones in spatially white Gaussian noise in the simulated room environment. The SNR is 10 dB.

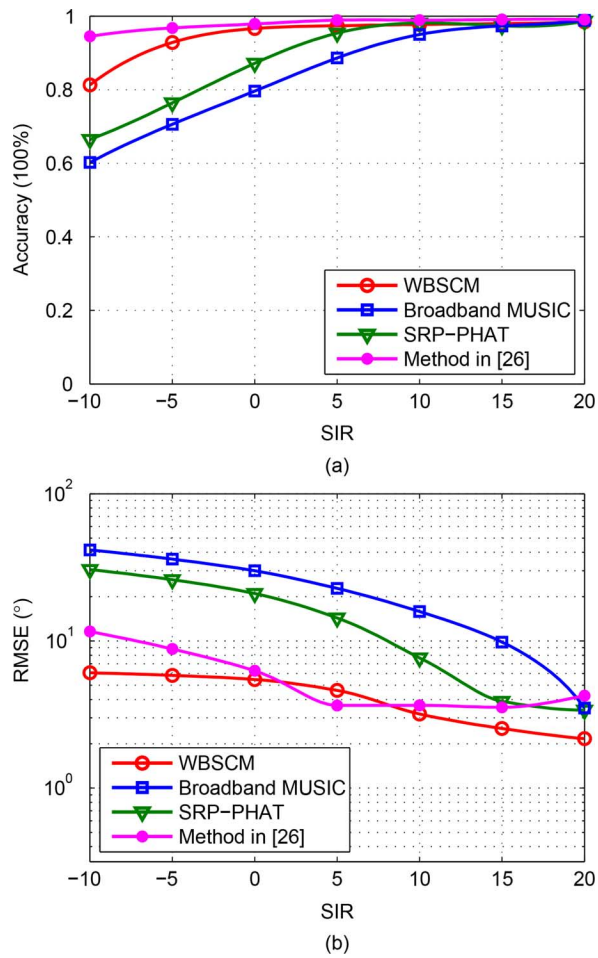


Fig. 10. (a) Estimation accuracy and (b) RMSE under different SIRs using car interior noise as the interference signal in the simulated room environment. Six microphones are used.

(car interior noise, F16 cockpit noise, white Gaussian noise, and speech babble noise). In each SIR and type of interference, different combinations of the speech DOA and interference DOA are considered, and the estimation results of all these considered scenarios are combined to achieve an overall evaluation. Same with the testing in spatially white Gaussian noise conditions, the speech DOA changes from $-90°$ to $90°$ with a step size of $20°$. For the interference, we simulate three different conditions, in which the interference DOAs are $-60°$, $-20°$, and $0°$ respectively. As the ULA is symmetric to the normal line, there is no need to consider the case that the interference appears on the opposite side. The received speech signal and interference signal are separately generated by the image-source method, then mixed together after being scaled to control the SIR. In all simulations of this subsection, six microphones are used.

The simulation results in directional car interior noise conditions are sketched in Fig. 10. We can observe that the proposed method is comparable with the method in [26] in high SIR scenarios on the estimation accuracy, while it is less robust than the method in [26] when the SIR is lower than -5 dB. Even so, its accuracy is still higher than 80% when the SIR is -10 dB, and much better than that of the SRP-PHAT and broadband MUSIC algorithm. Moreover, the proposed method can achieve the lowest RMSE in almost all SIRs evaluated.

The superiority of the proposed method over other comparison methods is more clearly shown in Figs. 11 and 12, where the F16 cockpit noise and white Gaussian noise are taken as interference signals. As we can see in Fig. 7, compared with the car interior noise, the F16 cockpit noise and white Gaussian noise have much broader frequency distributions. Especially, for the white Gaussian noise, it has the broadest frequency distribution with all bands having equal energy statistically. In the earlier part of this paper, we have discussed that if the interference signal has a flat frequency distribution, in the frequency domain, more speech bands will be polluted, making the DOA estimation problem more complicated. This is demonstrated by the observation that the performance of comparison methods generally degrades when the interference signal changes from the car interior noise to the white Gaussian noise, shown from Figs. 10 to 12. However, we can see that the proposed method degrades least. In the low SIR conditions, the proposed method exhibits much better performance than other methods.

As are illustrated in Figs. 11 and 12, the SRP-PHAT and broadband MUSIC algorithm almost totally break down when the SIR is lower than -5 dB. Even the SIR is higher than 0dB, generally, the performance of both methods is still not satisfactory. This is understandable. As are introduced in [12] and [11], the phase transform (PHAT) is adopted for computing the correlation functions between two signals in both methods. In the

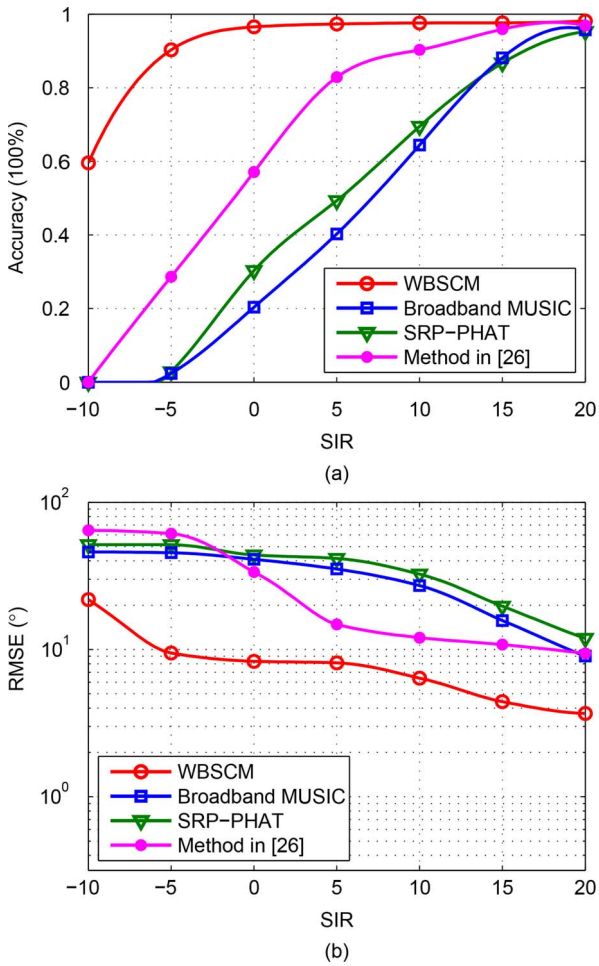


Fig. 11. (a) Estimation accuracy and (b) RMSE under different SIRs using F16 cockpit noise as the interference signal in the simulated room environment. Six microphones are used.

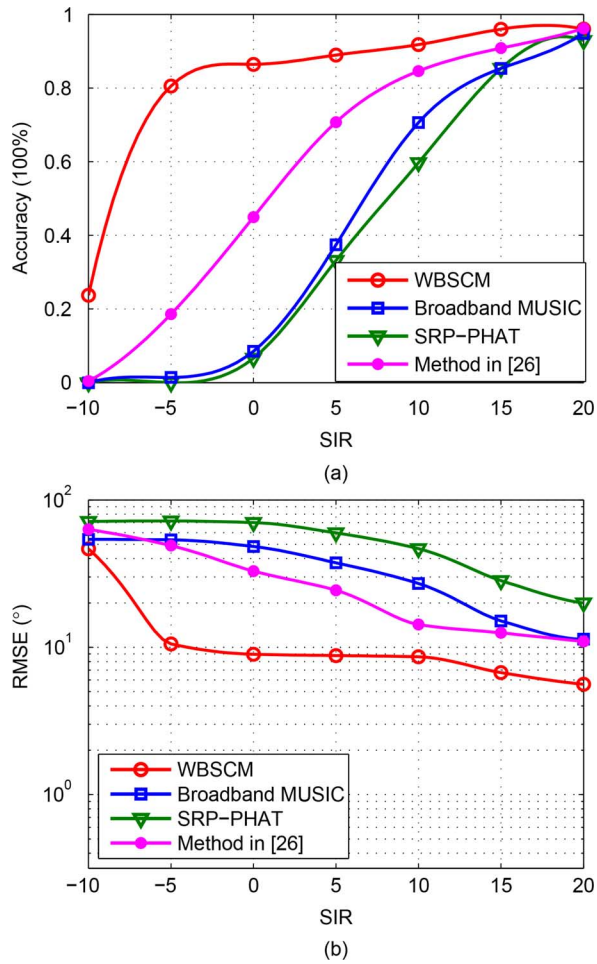


Fig. 12. (a) Estimation accuracy and (b) RMSE under different SIRs using white Gaussian noise as the interference signal in the simulated room environment. Six microphones are used.

frequency domain, this procedure discards the amplitude of the cross-spectrum in each frequency, which makes all frequency bins are treated with equal significance in the DOA cost function. As the speech signal occupies only a few frequency bins, when interference has a flat frequency distribution, most frequency bins are dominated by the interference rather than the speech. As a result, even though the noise level is lower than the speech, the DOA estimator may finally direct its global peak towards the noise source direction. The method in [26] improves the performance in low SIR conditions, by selecting the speech frequency bands for DOA estimation. Nevertheless, the proposed method, which utilizes the redundancy in BPDs to improve the robustness against the interference, produces the best results.

On the other hand, we notice that although the proposed method can work more robustly than other methods, it still suffers from performance degradation when the frequency distribution of interference gets broader, as are shown from Figs. 10 to 12. It is not surprising if comparing between the car interior noise case and F16 cockpit noise case. According to (6), the bispectrum reflects the interaction between different frequencies, then broad frequency distribution always results in broad bispectrum distribution, which eventually makes more speech bispectrum units polluted. However, the performance in white Gaussian noise cases is worse than that in the other two types of interference cases, this fact seems inconsistent with the theory that the HOS of white Gaussian noise is zero. It may be explained as follows. In practice, the bispectrum of white Gaussian noise is not exactly zero, and in low SIR conditions, the residual error in each bispectrum unit becomes also large. Moreover, as the theoretical "all-zeros" distribution is the most flat one, the bispectrum distribution of the residual error will be the broadest. Therefore, despite of the theoretical advantage of HOS over Gaussian noise, when the SIR is low, more speech units get affected than other interference cases, leading to less robust results compared with other types of interferences cases.

In the last set of experiments, the performance of different methods is evaluated in directional speech babble noises, and the results are shown in Fig. 13. As are illustrated in Figs. 7 and 6, compared with the other types of interference signals, the speech babble noise poses additional difficulties to the high performance DOA estimation, since its frequency distribution coincides much with that of the clean speech signal, and it is more non-stationary than other interference signals. Therefore, although the speech babble has less flat frequency distribution than the F16 cockpit noise, we can notice that, in general, the performance of different algorithms is even worse than that in F16 cockpit noise conditions, and it is only slightly better than that in white Gaussian noise conditions.

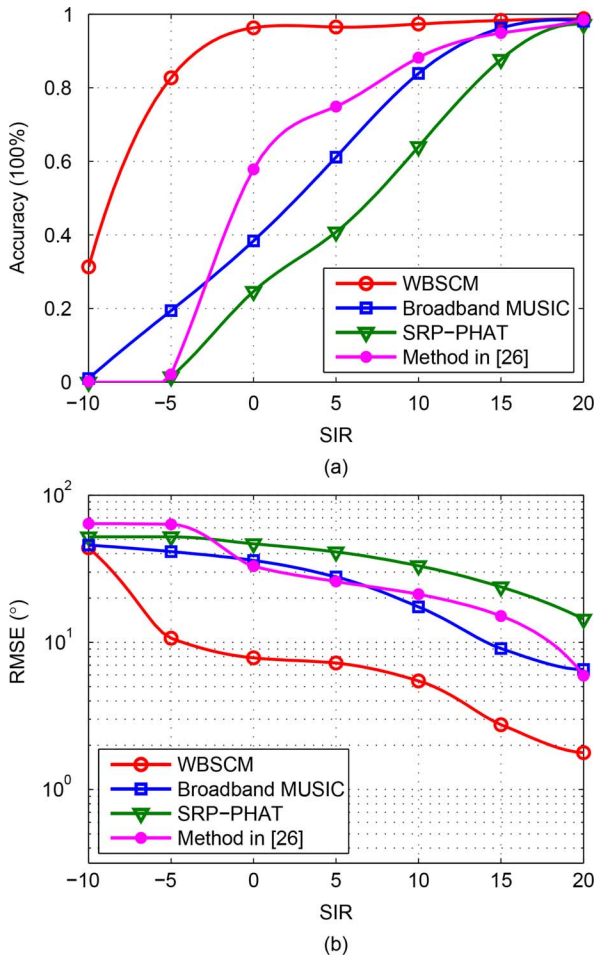


Fig. 13. (a) Estimation accuracy and (b) RMSE under different SIRs using speech babble noise as the interference signal in the simulated room environment. Six microphones are used.

E. Performance in Real Room Environments

In this subsection, by using the MIRD database, we evaluate the performance of different algorithms in real room conditions. During the short-time observation period, it can be assumed that the sound propagation from the source to microphone is linear and time-invariant, therefore the measured RIR entirely describes the real room environment [47]. Then different scenarios with various source DOAs in real room conditions can be reproduced by convolving the source signals with the RIRs. As the RIRs are measured with 48 kHz sampling rate, they are downsampled to 8 kHz (the sampling rate of source signals) before convolution. For the interference-existing conditions, the signal received by each microphone can be generated by mixing the received speech and interference signals, which are generated separately with the RIRs corresponding to speech and interference DOAs. However, for the spatially white noise conditions, as real noise recordings that are uncorrelated at different microphones are unavailable, we can not reproduce the this type of noisy testing data in real room environments. Therefore, in this subsection, we only test the performance in directional interference conditions.

Similar to the simulated interference-existing conditions, for the real room environments, various scenarios with different SIRs and types of interferences are reproduced. We consider three different SIRs, which are -5 dB, 5 dB and 15 dB, corresponding to the highly, moderately and slightly noisy environments respectively. Again, the car interior noise, F16 cockpit noise, white Gaussian noise, and speech babble noise are taken as interference signals. In each SIR and interference type, the speech source appears from -90° to 90° with step size as 30°, and the interference source is located at the direction of 15°, 45°, or 75°. We separately generate the received speech signal and interference signal using the real RIRs, and mix them at certain SIRs to produce the noisy observations. For each evaluated algorithm, in each SIR and interference type, the estimation results under different combinations of speech and interference DOAs are jointly analyzed for an overall evaluation.

The performance of different algorithms in real room environments are summarized in Fig. 14. From Fig. 14(a), (b), it can be seen that when the SIR $= -5$ dB, compared with the broadband MUSIC and SRP-PHAT algorithm, the proposed method can produce much better performance in all interference types considered. Although the method in [26] can work more reliably than the proposed method in car interior noise conditions, its performance is much lower than the proposed method in other interference types, especially when the speech babble noise exists. Comparing between the Fig. 14(a)$\sim$(f), we can observe that when the SIR gets higher, all algorithms can perform more robustly, nevertheless, in almost all scenarios, the proposed method achieves the best results. On the other hand, we can also notice that in each SIR, although different algorithms generally work better in car interior noise conditions, while less robustly in white Gaussian noise or speech babble noise conditions, the proposed method exhibits the least fluctuations in performance, which implies that it is less sensitive to the type of interference than the comparison methods.

VII. Conclusion

Estimating the DOA of the speech source is a challenging problem in noisy conditions. In this paper, we developed a new DOA estimation method which can perform robustly no matter the noises in different microphones are spatially white or directional. The proposed method is formulated in the bispectrum domain, and the core of the proposed method is the "WBSCM," which contains the spatial correlation information of multiple BPDs. As is HOS-based, the proposed method can exploit the vanishing property of HOS against the Gaussian noise. Moreover, by analyzing the BPD between the signals received by a pair of microphones, we showed that in the bispectrum domain, the speech DOA cue, which approximates to the BPD in speech-dominated units, is redundantly expressed, and the redundancy helps to improve the robustness of the algorithm. We proposed a decision-directed method to compute a set of bispectrum weights, which can be used to select the speech-dominated bispectrum units. By using the BPDs of multiple microphones and the computed bispectrum weights, we formulated a matrix called WBSCM. The WBSCM is a function of the hypothesized DOA, and exhibits a special property only when the hypothesized matches the true one. Finally, based on the eigenvalue analysis of the WBSCM, a new DOA estimator is further

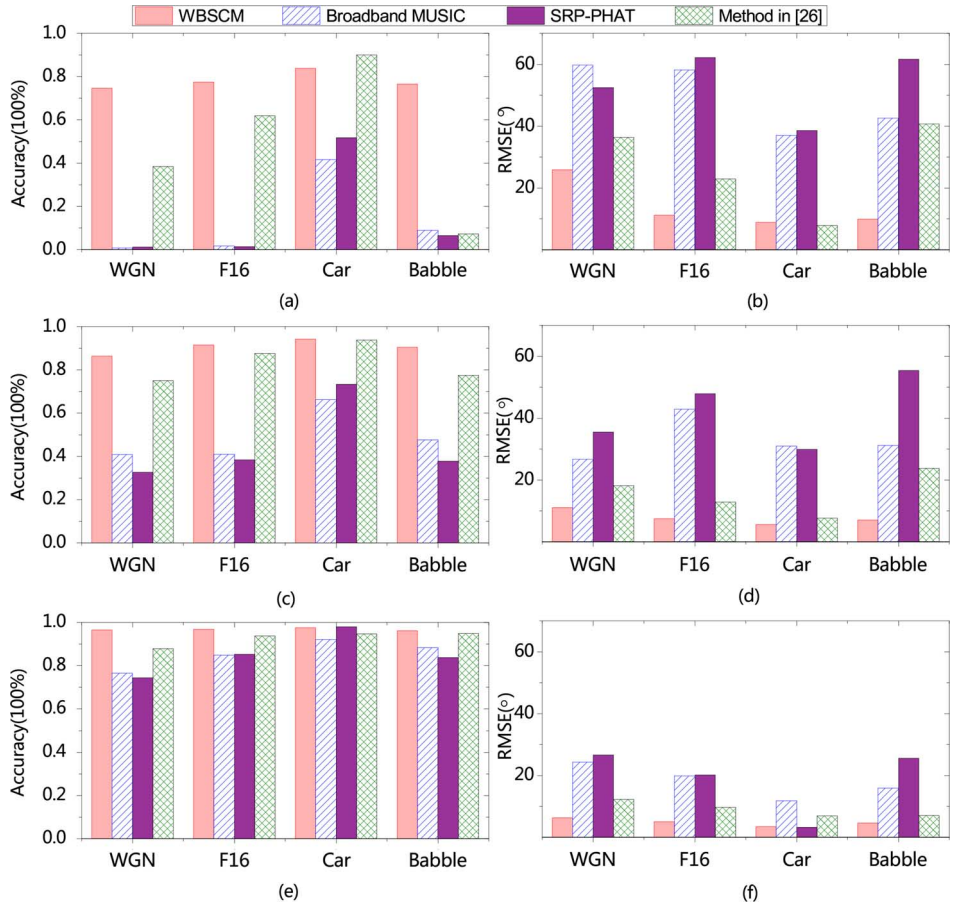


Fig. 14. Estimation accuracy and RMSE under different SIRs and types of interferences in real room conditions. The SIR is −5 dB for (a)(b), 5 dB for (c)(d), and 15 dB for (e) (f). The types of interference are white Gaussian noise (WGN), F16 cockpit noise (F16), car interior noise (Car), and speech babble noise (Babble), respectively.

developed. By conducting experiments under various kinds of noisy scenarios, we demonstrated the effectiveness of the proposed method.

Acknowledgment

We would like to thank the anonymous reviewers for their thorough work and valuable comments.

References

[1] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[2] S. Doclo and M. Moonen, "Design of broadband beamformers robust against gain and phase errors in the microphone array characteristics," *IEEE Trans. Signal Process.*, vol. 51, no. 10, pp. 2511–2526, Oct. 2003.

[3] Y. Huang, J. Benesty, and G. W. Elko, "Microphone arrays for video camera steering," in *Acoustic Signal Processing for Telecommunication*. New York, NY, USA: Springer, 2000, pp. 239–259.

[4] H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1997, vol. 1, pp. 187–190.

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

[6] J. Capon, "Maximum-likelihood spectral estimation," in *Nonlinear Methods of Spectral Analysis*. New York, NY, USA: Springer, 1983, pp. 155–179.

[7] R. Jeffers, K. L. Bell, and H. L. Van Trees, "Broadband passive range estimation using music," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2002, vol. 3, pp. 2921–2924.

[8] C. T. Ishi, O. Chatot, H. Ishiguro, and N. Hagita, "Evaluation of a music-based real-time sound localization of multiple sound sources in real noisy environments," in *Proc. IEEE/RSJ Int. Conf. Intell. Rob. Syst., IROS'09*, 2009, pp. 2027–2032.

[9] 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. Acoustics, Speech, Signal Process.*, vol. ASSP-33, no. 4, pp. 823–831, Aug. 1985.

[10] J. Krolik and D. Swingler, "Focused wide-band array processing by spatial resampling," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 2, pp. 356–360, Feb. 1990.

[11] J. P. Dmochowski, J. Benesty, and S. Affes, "Broadband music: Opportunities and challenges for multiple source localization," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, 2007, pp. 18–21.

[12] J. H. DiBiase, "A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays," Ph.D. dissertation, Brown Univ., Providence, RI, USA, 2000.

[13] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*. New York, NY, USA: Springer, 2010.

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

[15] M. S. Brandstein and H. F. Silverman, "A practical methodology for speech source localization with microphone arrays," *Comput. Speech Lang.*, vol. 11, no. 2, pp. 91–126, 1997.

[16] J. Benesty, M. M. Sondhi, and Y. Huang, *Springer Handbook of Speech Processing*. New York, NY, USA: Springer, 2008.

[17] C. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, no. 4, pp. 320–327, Aug. 1976.

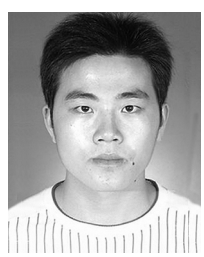
[18] J. P. Ianniello, "Time delay estimation via cross-correlation in the presence of large estimation errors," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-30, no. 6, pp. 998–1003, Dec. 1982.

[19] M. S. Brandstein, "Time-delay estimation of reverberated speech exploiting harmonic structure," *J. Acoust. Soc. Amer.*, vol. 105, no. 5, pp. 2914–2919, May 1999.
[20] B. Yegnanarayana, S. Prasanna, R. Duraiswami, and D. Zotkin, "Processing of reverberant speech for time-delay estimation," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 6, pp. 1110–1118, Nov. 2005.
[21] J. Benesty, J. Chen, and Y. Huang, "Time-delay estimation via linear interpolation and cross correlation," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 509–519, Sep. 2004.
[22] J. Chen, J. Benesty, and Y. Huang, "Robust time delay estimation exploiting redundancy among multiple microphones," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 549–557, Nov. 2003.
[23] J. Chen, J. Benesty, and Y. Huang, "Robust time delay estimation exploiting spatial correlation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'03)*, 2003, vol. 5, pp. V–481, IEEE.
[24] T. G. Dvorkind and S. Gannot, "Time difference of arrival estimation of speech source in a noisy and reverberant environment," *Signal Process.*, vol. 85, no. 1, pp. 177–204, 2005.
[25] T. Nishiura, S. Nakamura, and K. Shikano, "Talker localization in a real acoustic environment based on DOA estimation and statistical sound source identification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2002, vol. 1, pp. I–893.
[26] W. Xue, S. Liang, and W. Liu, "Interference robust DOA estimation of human speech by exploiting historical information and temporal correlation," in *Proc. INTERSPEECH*, 2013, pp. 2895–2899.
[27] P. Forster and C. L. Nikias, "Bearing estimation in the bispectrum domain," *IEEE Trans. Signal Process.*, vol. 39, no. 9, pp. 1994–2006, Sep. 1991.
[28] Z. Shi and F. W. Fairman, "A comprehensive approach to DOA estimation using higher-order statistics," *Circuits, Syst., Signal Process.*, vol. 17, no. 4, pp. 539–557, 1998.
[29] Z. Shi and F. W. Fairman, "DOA estimation via higher-order cumulants: A generalized approach," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1992, vol. 5, pp. 209–212.
[30] B. Porat and B. Friedlander, "Direction finding algorithms based on high-order statistics," *IEEE Trans. Signal Process.*, vol. 39, no. 9, pp. 2016–2024, Sep. 1991.
[31] N. Yuen and B. Friedlander, "DOA estimation in multipath: An approach using fourth-order cumulants," *IEEE Trans. Signal Process.*, vol. 45, no. 5, pp. 1253–1263, May 1997.
[32] W. Xue, S. Liang, and W. Liu, "DOA estimation of speech source in noisy environments with weighted spatial bispectrum correlation matrix," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 2014, pp. 2282–2286.
[33] W. Xue, S. Liang, and W. Liu, "Weighted spatial bispectrum correlation matrix for DOA estimation in the presence of interferences," in *Proc. INTERSPEECH*, 2014, pp. 2228–2232.
[34] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. New York, NY, USA: Springer, 2008, vol. 1.
[35] W. Collis, P. White, and J. Hammond, "Higher-order spectra: The bispectrum and trispectrum," *Mech. Syst. Signal Process.*, vol. 12, no. 3, pp. 375–394, 1998.
[36] J. W. Tukey, *The Collected Works of John W. Tukey*. New York, NY, USA: Taylor & Francis, 1994, vol. 1.
[37] D. Brillinger, "The identification of polynomial systems by means of higher order spectra," *J. Sound Vibr.*, vol. 12, no. 3, pp. 301–313, 1970.
[38] D. R. Brillinger, "An introduction to polyspectra," *Ann. Math. Statist.*, pp. 1351–1374, 1965.
[39] C. L. Nikias and M. R. Raghuveer, "Bispectrum estimation: A digital signal processing framework," *Proc. IEEE*, vol. 75, no. 7, pp. 869–891, Jul. 1987.
[40] C. L. Nikias and J. M. Mendel, "Signal processing with higher-order spectra," *IEEE Signal Process. Mag.*, vol. 10, no. 3, pp. 10–37, Jul. 1993.
[41] J. Fackrell and S. McLaughlin, "The higher-order statistics of speech signals," in *IEE Colloquium Tech. Speech Process. Their Applicat.*, 1994, pp. 7–11.
[42] C. L. Nikias and R. Pan, "Time delay estimation in unknown gaussian spatially correlated noise," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 11, pp. 1706–1714, Nov. 1988.
[43] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.
[44] D. Malah, R. Cox, and A. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1999, vol. 2, pp. 789–792.
[45] C. Breithaupt, T. Gerkmann, and R. Martin, "Cepstral smoothing of spectral filter gains for speech enhancement without musical noise," *IEEE Signal Process. Lett.*, vol. 14, no. 12, pp. 1036–1039, Dec. 2007.
[46] E. A. Lehmann and A. M. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *J. Acoust. Soc. Amer.*, vol. 124, p. 269, 2008.
[47] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *Proc. 14th Int. Workshop Acoust. Signal Enhance. (IWAENC)*, 2014, pp. 313–317.
[48] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. Noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, 1993.

Wei Xue was born in 1989. He received his Bachelor degree in automation from Huazhong University of Science and Technology, Wuhan, China, in 2010. He is currently a Ph.D. student in the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His current research interests are in sound source localization, speech enhancement, and speech separation.

Wenju Liu was born in 1960. He received the B.S. and M.S. degrees in mathematics from Beijing University and Beijing University of Post and Telecommunication, and the Ph.D. degree in computer applications from Tsinghua University, Beijing, China, in 1983, 1989, and 1993, respectively. Currently, he is a Research Professor at the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include speech recognition, speech synthesis, speaker recognition, key words spotting, computational auditory scene analysis, speech enhancement, noise reduction, etc.

Dr. Wenju Liu is a member of neural network committee of China and the signal processing society of the IEEE. He is an editorial board member of *Journal of Computer Science Application* as well as a reviewer of numerous academic journals such as the IEEE TRANSACTION ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, *Computer Speech and Language*, *Speech Communication*, *Cognitive Computation*, etc.



Shan Liang was born in 1987. He received the Ph.D. degree at the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2013 and the B.Sc. degree in computer science and technology from the XiDian University, Xi'an, China, in 2008. Currently, he is a Research Assistant In NLPR.

His current research concentrates on computational auditory scene analysis, blind source separation, and speech enhancement areas.