ROBUST SPEAKER DOA ESTIMATION WITH SINGLE AVS IN BISPECTRUM DOMAIN

*Y. H. Jin, Y. X. Zou**

ADSPLAB, School of ECE, Peking University, Shenzhen, 518055, China
*Corresponding author: zouyx@pkusz.edu.cn

ABSTRACT

For mobile speech application, speaker DOA estimation accuracy, interference robustness and compact physical size are three key factors. Considering the size, we utilized acoustic vector sensor (AVS) and proposed a DOA estimation algorithm previously [1], offering high accuracy with larger-than-15dB SNR but is deteriorated by non-speech interferences (NSI). This paper develops a robust speaker DOA estimation algorithm. It is achieved by deriving the inter-sensor data ratio model of an AVS in bispectrum domain (BISDR) and exploring the favorable properties of bispectrum, such as zero value of Gaussian process and different distribution of speech and NSI. Specifically, a reliable bispectrum mask is generated to guarantee that the speaker DOA cues, derived from BISDR, are robust to NSI in terms of speech sparsity and large bispectrum amplitude of the captured signals. Intensive experiments demonstrate an improved performance of our proposed algorithm under various NSI conditions even when SIR is smaller than 0dB.

Index Terms— Direction of arrival estimation, acoustic vector sensor, bispectrum inter-sensor data ratio, interference

1. INTRODUCTION

Direction of arrival (DOA) estimation in an acoustic environment has a wide range of applications such as video conferencing and intelligent robots for identifying the speech source localization swiftly and accurately [2]. For the applications against background noises, conventional approaches often utilize an array of omni-directional microphones and DOA estimation is achieved by exploiting phase-delay information between the microphones [3]. However, conventional arrays often require a large aperture, which presents limits in space-constrained applications.

By comparison, AVS is more attractive for mobile speech applications [4, 5] as it can provide more information with a smaller size and no spatial aliasing limitation [6]. The AVS has been employed to improve the DOA estimation performance by different methods such as exploiting beamforming and subspace methods [7–12].

For multisource DOA estimation, in our previous work [1], a single AVS based algorithm has been developed with the definition of inter-sensor data ratio (ISDR) in the spectrum domain. By exploring the harmonic structure of speech, the time-frequency points with high local signal-to-noise ratio (HLSNR-TFPs) were extracted using the Sinusoidal tracks extraction (SinTrE) method. Then the elevation and azimuth angles are easy to be estimated by kernel density estimation (KDE) on the ISDR at the HLSNR-TFPs. However, our experiments showed that the HLSNR-TFPs extraction method proposed in [1] was not robust to noise and NSI, which degrades the DOA estimation performance.

This work is partially supported by National Natural Science Foundation of China (No: 61271309).

As the literature shows, NSI should be considered in the practical applications, such as the air-condition noise for the audition system of the service robot and machine noise for the automatic camera steering. However, there are very few research outcomes reported in regard to them.

This paper intends to focus on the robust DOA estimation when NSI exists. The AVS is compact in size so is used to capture signals. Our basic idea is to find an effective measure to extract the DOA information of a speaker while at the same time suppress the unwanted additive background noise and NSI. Bispectrum is a kind of high order statistics (HOS) of signal defined in terms of high order cumulants of the random process. The motivation of choosing to work with bispectrum lies in following aspects: 1) The HOS of the Gaussian process is always zero [13], which has been used for developing several DOA estimation algorithms robust to Gaussian noise with narrowband signals [14–18]. And a recent study adopted HOS to overcome the effect of spatially colored Gaussian noise for speech DOA estimation [19]. 2) The bispectra of speech and non-speech signal developed in this study are different, which means most of speech DOA cues will not be suppressed by noise and NSI mentioned above. As a result, even though [19] is not capable to deal with the non-Gaussian interferences directly since their bispectra are non-zero, we can explore redundancy of the speech DOA cues in bispectrum domain to reduce the adverse impact of the NSI.

According to the discussions above, in the following paper, we firstly introduce the data model of an AVS and derive its bispectrum expression. Inspired by the previous DOA estimation method developed in spectrum domain [1], the inter-sensor data ratios model in the bispectrum domain termed as BISDR is then derived. Through the analysis of its property, the DOA estimation problem is formulated as extracting the reliable DOA-related information from the BISDR at certain frequency points with high local signal-to-interference ratio (HLSIR-FPs). The bivariate KDE technique is employed as the clustering algorithm to estimate DOA cues and thereafter the DOA is estimated. It is obvious that one of the key steps for the proposed DOA estimation method is to determine the HLSIR-FPs. In our study, a reliable bispectrum mask is derived for it. Intensive experiments under various NSI conditions and recording data have been carried out to demonstrate the effectiveness and robustness of the proposed DOA estimation method.

Notation: Throughout the paper, superscripts T and $*$ represent the matrix or vector transpose and convolution, respectively.

2. DATA MODELS

2.1. AVS Data Models

Generally each AVS unit consists of an omnidirectional sensor (o-sensor) and three orthogonally oriented directional sensors (named as u-, v-, w-sensor, respectively). Supposing there is one speech signal $s(k)$ impinging upon the AVS unit with the DOA of (θ_s, ϕ_s)

in which the elevation angle $\theta_s \in (0^\circ, 180^\circ)$ and the azimuth angle $\phi_s \in [0^\circ, 360^\circ)$, its associated manifold vector is given by [1]

$$\boldsymbol{a}(\theta_s, \phi_s) \equiv [u_s, v_s, w_s, 1]^T, \boldsymbol{a} \in R^{4\times 1} \quad (1)$$

where the elements u_s, v_s, and w_s are the x-,y-, and z-axis direction cosines, respectively. They can be determined according to the unit geometry, which is derived as follows:

$$u_s = \sin\theta_s \cos\phi_s, v_s = \sin\theta_s \sin\phi_s, w_s = \cos\theta_s \quad (2)$$

Then, the data captured by AVS at time k can be expressed as

$$\boldsymbol{x}(k) = \boldsymbol{a}(\theta_s, \phi_s)s(k) * h_s(k) + \sum_{i=1}^{L} \boldsymbol{a}(\theta_{ri}, \phi_{ri})r_i(k) * h_{ri}(k) + \boldsymbol{n}(k) \quad (3)$$

where $\boldsymbol{x}(k) = [x_u(k), x_v(k), x_w(k), x_o(k)]^T$ represents the output of the u-, v-, w-, and o-sensor, respectively; $s(k)$ is the speech signal with DOA (θ_s, ϕ_s) and the room impulse response $h_s(k)$; $r_i(k)$ is the i^{th} NSI signal with DOA (θ_{ri}, ϕ_{ri}) and the room impulse response $h_{ri}(k)$, which is assumed uncorrelated to the speech signal; And $\boldsymbol{n}(k) = [n_u(k), n_v(k), n_w(k), n_o(k)]^T$ denotes the zero-mean additive white Gaussian noise (AWGN) at the u-, v-, w-, and o-sensor, respectively.

In order to facilitate the discussion, we just assume $L = 1$. Then the data model in (3) can be further expressed as:

$$x_u(k) = u_s s(k) * h_s(k) + u_r(k)r(k) * h_r(k) + n_u(k) \quad (4)$$

$$x_v(k) = v_s s(k) * h_s(k) + v_r(k)r(k) * h_r(k) + n_v(k) \quad (5)$$

$$x_w(k) = w_s s(k) * h_s(k) + w_r(k)r(k) * h_r(k) + n_w(k) \quad (6)$$

$$x_o(k) = s(k) * h_s(k) + r(k) * h_r(k) + n_o(k) \quad (7)$$

2.2. Bispectrum Domain Representation

According to the derivation in [20], the following cross-relationships between $x_o(k)$ and $x_j(k)$ (for j refers to u, v, w, o) hold in the third moment domain for the data given by (4)-(7). We have

$$R_{x_o x_j x_o}(\tau, \rho) = E\{x_o(k)x_j(k+\tau)x_o(k+\rho)\} \quad (8)$$

With the assumption that s, r, and n are uncorrelated with each other, and $s_h(k) = s(k) * h_s(k)$, $r_h(k) = r(k) * h_r(k)$, taking $j = u$ as an example, substituting (4) and (7) into (8) gives

$$\begin{aligned} R_{x_o x_u x_o}(\tau, \rho) = &E\{s_h(k)u_s s_h(k+\tau)s_h(k+\rho)\} \\ &+ E\{r_h(k)u_r r_h(k+\tau)r_h(k+\rho)\} \\ &+ E\{n_o(k)n_u(k+\tau)n_o(k+\rho)\} \end{aligned} \quad (9)$$

It is noted that $E\{n_o(k)n_u(k+\tau)n_o(k+\rho)\} = 0$ [13], then (9) can be simplified as

$$\begin{aligned} R_{x_o x_u x_o}(\tau, \rho) = &u_s E\{s_h(k)s_h(k+\tau)s_h(k+\rho)\} \\ &+ u_r E\{r_h(k)r_h(k+\tau)r_h(k+\rho)\} \end{aligned} \quad (10)$$

The bispectrum is, by definition, the Fourier transform of the third moment sequence viz. [21–23]. With the linearity of the Fourier transform, from (10), we get

$$\begin{aligned} B_{x_o x_u x_o}(\Omega_1, \Omega_2) = &FT[R_{x_o x_u x_o}(\tau, \rho)] \\ = &FT[u_s E\{s_h(k)s_h(k+\tau)s_h(k+\rho)\}] \\ &+ FT[u_r E\{r_h(k)r_h(k+\tau)r_h(k+\rho)\}] \end{aligned} \quad (11)$$

where the $FT[\cdot]$ denotes the 2-D Fourier transform operation. With the definition, (11) can be further expressed as

$$B_{x_o x_u x_o}(\Omega_1, \Omega_2) = u_s B_{s_h s_h s_h}(\Omega_1, \Omega_2) + u_r B_{r_h r_h r_h}(\Omega_1, \Omega_2) \quad (12)$$

With the same procedure, we can derive the followings:

$$B_{x_o x_v x_o}(\Omega_1, \Omega_2) = v_s B_{s_h s_h s_h}(\Omega_1, \Omega_2) + v_r B_{r_h r_h r_h}(\Omega_1, \Omega_2) \quad (13)$$

$$B_{x_o x_w x_o}(\Omega_1, \Omega_2) = w_s B_{s_h s_h s_h}(\Omega_1, \Omega_2) + w_r B_{r_h r_h r_h}(\Omega_1, \Omega_2) \quad (14)$$

$$B_{x_o x_o x_o}(\Omega_1, \Omega_2) = B_{s_h s_h s_h}(\Omega_1, \Omega_2) + B_{r_h r_h r_h}(\Omega_1, \Omega_2) \quad (15)$$

Analyzing (12)-(15), we can see that the first terms of (12)-(15) are only related to the speech source, and the second terms are only related to the directional non-speech interferences. Our target is to get the DOA information of the speech signal embedded in the first terms being able to suppress the adverse impacts of the second terms.

3. PROPOSED DOA ESTIMATION METHOD

3.1. Bispectrum Inter-Sensor Data Ratios (BISDR)

Following the idea of [1], in this subsection, we define the BISDR of the AVS as follows:

$$I_{uo}(\Omega_1, \Omega_2) = B_{x_o x_u x_o}(\Omega_1, \Omega_2)/B_{x_o x_o x_o}(\Omega_1, \Omega_2) \quad (16)$$

$$I_{vo}(\Omega_1, \Omega_2) = B_{x_o x_v x_o}(\Omega_1, \Omega_2)/B_{x_o x_o x_o}(\Omega_1, \Omega_2) \quad (17)$$

$$I_{wo}(\Omega_1, \Omega_2) = B_{x_o x_w x_o}(\Omega_1, \Omega_2)/B_{x_o x_o x_o}(\Omega_1, \Omega_2) \quad (18)$$

where $I_{uo}(\Omega_1, \Omega_2)$, $I_{vo}(\Omega_1, \Omega_2)$, and $I_{wo}(\Omega_1, \Omega_2)$ are termed as the BISDR between u- and o-sensor, v- and o-sensor, w- and o-sensor, respectively. Taking $I_{uo}(\Omega_1, \Omega_2)$ as an example, substituting

$$\begin{aligned} I_{uo}(\Omega_1, \Omega_2) = &\frac{u_s B_{s_h s_h s_h}(\Omega_1, \Omega_2) + u_r B_{r_h r_h r_h}(\Omega_1, \Omega_2)}{B_{s_h s_h s_h}(\Omega_1, \Omega_2) + B_{r_h r_h r_h}(\Omega_1, \Omega_2)} \\ = &\frac{u_s[B_{s_h s_h s_h}(\Omega_1, \Omega_2) + B_{r_h r_h r_h}(\Omega_1, \Omega_2)]}{B_{s_h s_h s_h}(\Omega_1, \Omega_2) + B_{r_h r_h r_h}(\Omega_1, \Omega_2)} \\ &+ \frac{u_r B_{r_h r_h r_h}(\Omega_1, \Omega_2) - u_s B_{r_h r_h r_h}(\Omega_1, \Omega_2)}{B_{s_h s_h s_h}(\Omega_1, \Omega_2) + B_{r_h r_h r_h}(\Omega_1, \Omega_2)} \end{aligned} \quad (19)$$

To simplify (19), rewrite it as follows:

$$I_{uo}(\Omega_1, \Omega_2) = u_s + \varepsilon_u(\Omega_1, \Omega_2) \quad (20)$$

where the second term is given by

$$\begin{aligned} \varepsilon_u(\Omega_1, \Omega_2) = &\frac{u_r B_{r_h r_h r_h}(\Omega_1, \Omega_2) - u_s B_{r_h r_h r_h}(\Omega_1, \Omega_2)}{B_{s_h s_h s_h}(\Omega_1, \Omega_2) + B_{r_h r_h r_h}(\Omega_1, \Omega_2)} \\ = &\frac{u_r - u_s}{1 + B_{s_h s_h s_h}(\Omega_1, \Omega_2)/B_{r_h r_h r_h}(\Omega_1, \Omega_2)} \end{aligned} \quad (21)$$

Similarly, $I_{vo}(\Omega_1, \Omega_2)$ and $I_{wo}(\Omega_1, \Omega_2)$ can be modeled as:

$$I_{vo}(\Omega_1, \Omega_2) = v_s + \varepsilon_v(\Omega_1, \Omega_2) \quad (22)$$

$$I_{wo}(\Omega_1, \Omega_2) = w_s + \varepsilon_w(\Omega_1, \Omega_2) \quad (23)$$

where

$$\varepsilon_v(\Omega_1, \Omega_2) = \frac{v_r - v_s}{1 + B_{s_h s_h s_h}(\Omega_1, \Omega_2)/B_{r_h r_h r_h}(\Omega_1, \Omega_2)} \quad (24)$$

$$\varepsilon_w(\Omega_1, \Omega_2) = \frac{w_r - w_s}{1 + B_{s_h s_h s_h}(\Omega_1, \Omega_2)/B_{r_h r_h r_h}(\Omega_1, \Omega_2)} \quad (25)$$

The compact expression of the BISDR can be written as:

$$\boldsymbol{I}(\Omega_1, \Omega_2) = \boldsymbol{b}(\theta_s, \phi_s) + \boldsymbol{\varepsilon}(\Omega_1, \Omega_2) \quad (26)$$

where

$$\boldsymbol{I}(\Omega_1, \Omega_2) = [I_{uo}(\Omega_1, \Omega_2), I_{vo}(\Omega_1, \Omega_2), I_{wo}(\Omega_1, \Omega_2)]^T \quad (27)$$

$$\boldsymbol{b}(\theta_s, \phi_s) = [u_s, v_s, w_s]^T \quad (28)$$

$$\boldsymbol{\varepsilon}(\Omega_1, \Omega_2) = [\varepsilon_u(\Omega_1, \Omega_2), \varepsilon_v(\Omega_1, \Omega_2), \varepsilon_w(\Omega_1, \Omega_2)]^T \quad (29)$$

In (26), there are two terms in the BISDR. Obviously, the first term $\boldsymbol{b}(\theta_s, \phi_s)$ is only related to the DOA of the speech source, which we call "speech DOA cue". It is clear that if we are able to find the FPs so that the second term in (26) approaches zero vector, then speech DOA cue can be perfectly estimated by (26) since the BISDR $\boldsymbol{I}(\Omega_1, \Omega_2)$ are available.

From (21), (24) and (25), we can see that if point $(\Omega_{1g}, \Omega_{2g})$ satisfies $B_{s_h s_h s_h}(\Omega_{1g}, \Omega_{2g}) \gg B_{r_h r_h r_h}(\Omega_{1g}, \Omega_{2g})$, each element of $\boldsymbol{\varepsilon}(\Omega_{1g}, \Omega_{2g})$ approximates 0. Then, we name these points as HLSIR-FPs. Judged from the above discussion, it is also the speech-

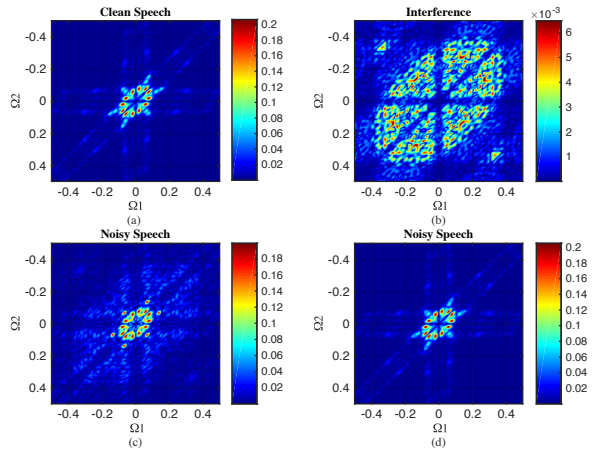

Fig. 1. Example of the bispectrum amplitude of the signal of o-sensor. Interference: hfchannel noise, SIR = -5dB in (c) and SIR = 10dB in (d). AWGN, SNR = 10dB.

dominated point where the real DOA cues are close to the BISDR. When $L > 1$, we can also get the similar results.

In this case, we should have some detailed analysis of the properties of the bispectra of speech signal and non-speech signal. As an example, we visualize the bispectra plots of the speech and NSI at different SIR conditions in Fig.1. We have the following observations: 1) Comparing Fig.(a) and (b), the spread of bispetra for speech and non-speech signals are different. 2) Comparing Fig.(a) and (c), we can clearly see that the pattern of the speech bispectrum can be observed mostly in the large-amplitude areas even when SIR=-5dB. 3) Comparing Fig.(a), (c) and (d), it is clear that with the increase of SIR, the adverse impact of directional NSI on the bispectrum of the speech reduces. 4) Obviously, from Fig.(c) and (d), we are confident that we are able to find some FPs where the speech information dominates, which will lead us estimate the speech DOA cues from (26) properly. In other words, in the speech-dominated FPs, the speech DOA cue can be approximated by the BISDR. In the next subsection, we will introduce the method to determine the speech-dominated FPs.

3.2. Bispectrum Mask Estimation

From the analysis above, the idea to determine the speech dominated FPs is very straightforward since the NSI is not able to corrupt the bispectrum of the speech signal (referring to Fig.1.(c), it is clear to see the speech bispectrum pattern) at the FPs with large amplitude. This means we can directly threshold the bispectrum of o-sensor to extract a so-called speech dominated bispectrum mask and the associated HLSIR-FPs can be obtained. These HLSIR-FPs then are applied to compute the corresponding BISDR.

Specifically, in our study, the speech-dominated bispectrum mask $m(\Omega_1, \Omega_2)$ can be determined using o-sensor signal as:

$$m(\Omega_1,\Omega_2)=\begin{cases}1 & |B_{x_ox_ox_o}(\Omega_1,\Omega_2)|>\xi\max(|B_{x_ox_ox_o}(\Omega_1,\Omega_2)|)\\ 0 & \text{otherwise}\end{cases} \quad (30)$$

where $|\cdot|$ is the amplitude operator and the threshold ξ is set as 0.7 by empirical results, which will be described in section 4. With this bispectrum mask, the BISDR at the speech-dominated FPs can be computed as follows:

$$\widehat{\boldsymbol{I}}(\Omega_1,\Omega_2) = m(\Omega_1,\Omega_2)\boldsymbol{I}(\Omega_1,\Omega_2) \quad (31)$$

where $\widehat{\boldsymbol{I}}(\Omega_1, \Omega_2)$ is the masked BISDR. Comparing (31) with (26), the majority of the second term $\boldsymbol{\varepsilon}(\Omega_1, \Omega_2)$ in (26) will be removed with mainly the real speech DOA cue $\boldsymbol{b}(\theta_s, \phi_s)$ left.

3.3. DOA Estimation Algorithm

From the description above, it is obvious that the BISDR can be viewed as the random variables in the bispectrum domain with mean of u_x, v_s, and w_s, respectively. Specifically, the DOA estimation task is to estimate the cluster centers at $(\hat{u}_s, \hat{v}_s, \hat{w}_s)$ by clustering the BISDR corresponding to all HLSIR-FPs. To achieve an effective and robust clustering result, KDE method is adopted in our study [24]. With the clustering result $(\hat{u}_s, \hat{v}_s, \hat{w}_s)$, according to (2), the estimated DOA $(\hat{\theta}_s, \hat{\phi}_s)$ can be calculated as

$$\hat{\theta}_s = \cos^{-1}\hat{w}_s, \hat{\phi}_s = \tan^{-1}(\hat{v}_s/\hat{u}_s) \quad (32)$$

To simplify the notation in the following context, the proposed DOA estimation algorithm is termed as the **AVS-BISDR** algorithm, which is developed under the cluster of BISDR data using single AVS. The AVS-BISDR algorithm is summarized as follows:

1) Segment the AVS output data.
2) Calculate the bispectrum of the four sensors by (12)-(15).
3) Calculate the BISDR between sensors by (16)-(18).
4) Get the bispectrum mask by (30) and add it on the BISDR.
5) Estimate the DOA via (31) by the clustering result derived using KDE [24].

4. EXPERIMENTAL RESULTS

In this section, several experiments are carried out to evaluate the performance of our proposed AVS-BISDR algorithm under different conditions. The GMDA-Laplace algorithm [6] and AVS-ISDR algorithm [1] are taken as the comparison methods.

The simulation experimental settings are as follows: the speech signal is of 3 seconds and sampled at 8kHz. One NSI is set at $(60^\circ, 75^\circ)$ and no reverberation is considered. The type of the NSI can be white Gaussian noise, hfchannel noise, pink noise, or factory noise taken from Noisex92 [25]. In addition, the AWGN is taken to simulate a more adverse environment. For processing the signals, the frame size is set to be 256 samples with 60% overlap. It is noted that, for the GMDA-Laplace algorithm, following the setup in [6], the DOA estimation results are obtained by running two times of the algorithm since originally the GMDA-Laplace algorithm only use two microphones placed along one axis with $8cm$ spacing.

The root mean squared error (RMSE) averaging over one hundred independent trials is taken as the performance metric, which is defined as $RMSE = 0.5\sqrt{\sum_{l=1}^{100}((\hat{\theta}_l - \theta)^2 + (\hat{\phi}_l - \phi)^2)/100}$, where $\hat{\theta}_l$ and $\hat{\phi}_l$ are respectively the estimated angles of the target speaker angles θ and ϕ on the l^{th} trial.

In the first experiment, we aim to evaluate the impact of choosing different thresholds of the mask on the performance of our proposed algorithm under different SIR conditions since the threshold is an important parameter for it. The speech source is set at $(60^\circ, 45^\circ)$. The results are shown in Fig. 2. It is noted that the RMSE reduces with the increase of SIR. Besides, when $\xi > 0.5$, the RMSE is not sensitive to the choice of ξ. The optimal ξ can be selected as 0.7, which gives the best results under most SIR conditions. Indirectly, these results further validate the effectiveness of our proposed bispectrum mask for HLSIR-FPs extraction.

The second experiment is conducted to evaluate the sensitivity of the proposed DOA algorithm over different azimuth angles. To visualization purpose, we fix $\theta_s = 60^\circ$. The experimental results are

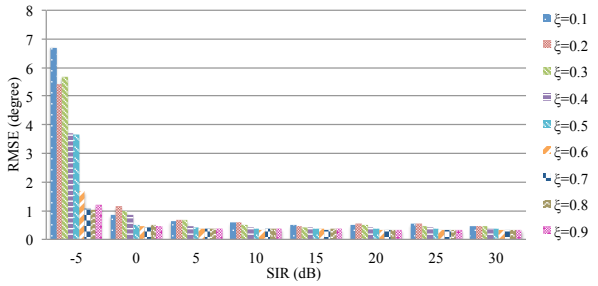

Fig. 2. RMSE versus SIR levels using different mask thresholds. Interference: factory noise. AWGN, SNR = 10dB.

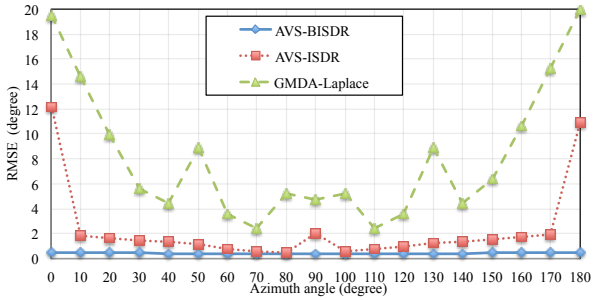

Fig. 3. RMSE versus different source azimuth angles. Interference: hfchannel noise, SIR = 5dB. AWGN, SNR=10dB.

shown in Fig. 3. We are encouraged to see that our proposed AVS-BISDR algorithm outperforms the comparison algorithms, where the RMSE values are closed to 0° for all angles. The AVS-ISDR algorithm performs the second best. It is noted that the RMSE values are below 2° except the values at three special angles ($\phi_s = 0^\circ$, 90°, and 180°). For the GMDA-Laplace algorithm, we can see the significant impact of the NSI. The RMSEs of the GMDA-Laplace algorithm range from 3° to 20°. It is obvious that the GMDA-Laplace algorithm performs worst at two special angles ($\phi_s = 0^\circ$ and 180°).

The third experiment aims at evaluating the robustness of the proposed AVS-BISDR under different input SIRs and types of NSI. The speech source is set at $(60^\circ, 45^\circ)$. Experimental results are presented in Fig. 4. We can see that the RMSEs of the proposed method under all the four interferences are constantly close to 0° even when the SIR is less than 0dB. For the other two algorithms, they both suffer a severe decline of the DOA estimation performance on the impact of the strong interferences. This verifies our proposed algorithm is more effective and robust under NSI.

In the fourth experiment, the behavior of the AVS-BISDR under different reverberation levels is evaluated. The experimental setup is as follows: A rectangular room with size $10m \times 5m \times 4m$ is modeled in the experiment by the image method [26]. Five different reverberation time (RT_{60}) conditions are considered. The speech source is located at DOA of $(60^\circ, 45^\circ)$. It is seen in Fig. 5 that the curve of the proposed method is approximately constant and keep a lower RMSE than that of AVS-ISDR for all RT_{60} conditions. This indicates that our proposed method is not sensitive to the room reverberation, which is a very favorable property since the performance of many existing DOA estimation algorithms, such as GMDA-Laplace, degrades when heavy room reverberation exists.

The last experiment is conducted to evaluate the performance of the proposed AVS-BISDR algorithm in a real scenario using the AVS data capturing system developed by ADSPLAB [1]. Five different groups of DOA are estimated respectively in Table 1. We are happy

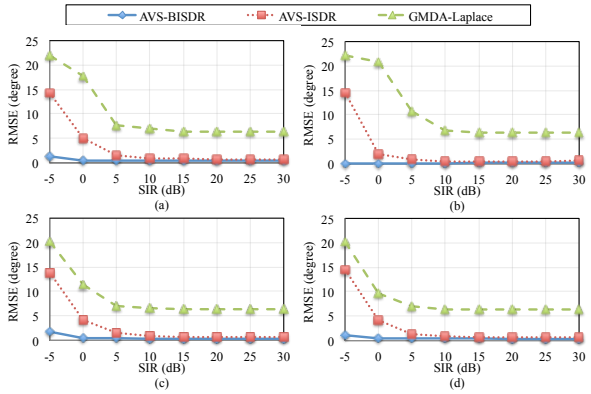

Fig. 4. RMSE versus different SIR levels and interference signals as: white Gaussian noise (a), hfchannel noise (b), pink noise (c), and factory noise (d). AWGN, SNR=10dB.

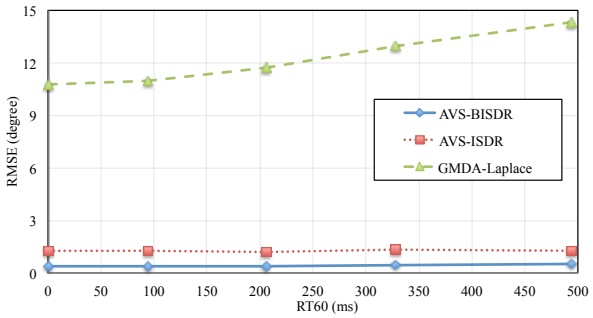

Fig. 5. RMSE versus different RT_{60}. Interference: hfchannel noise, SIR = 5dB. AWGN, SNR=10dB.

Table 1. DOA estimation resluts in a real scenario

True DOA	θ	90°	90°	90°	90°	90°
	ϕ	0°	45°	90°	135°	180°
AVS-BISDR	θ	93.57°	93.52°	90.09°	91.92°	89.42°
	ϕ	1.04°	43.56°	90.36°	132.61°	179.61°

to see that the DOA estimation errors of the proposed AVS-BISDR algorithm are less than 5° in each group in the real scenario.

5. CONCLUSIONS

In this paper, a novel interference robust DOA estimation method for speech source (termed as AVS-BISDR) has been developed in the bispectrum domain using single AVS. The key idea of deriving AVS-BISDR algorithm is to exploit the speech HOS spatial location information embedded in the AVS, and extract the HLSIR-FPs for the DOA estimation of the speech source with the bispectrum mask. Extensive experiments have been conducted to evaluate the performance of the proposed AVS-BISDR under different SIR levels and interference conditions. Results validate the superior performance of our proposed AVS-BISDR using simulated and real captured data. It is found that AVS-BISDR is able to obtain high DOA estimation accuracy even under strong interferences. Our future work will focus on the multisource DOA estimation in non-stationary interference situations with strong noises.

6. REFERENCES

[1] Yue Xian Zou, Wei Shi, Bo Li, Christian H Ritz, Muawiyath Shujau, and Jiangtao Xi, "Multisource doa estimation based on time-frequency sparsity and joint inter-sensor data ratio with single acoustic vector sensor," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 4011–4015.

[2] Flavio Ribeiro, Cha Zhang, Dinei Florencio, Demba Elimane Ba, et al., "Using reverberation to improve range and elevation discrimination for small array sound source localization," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 7, pp. 1781–1792, 2010.

[3] Jacob Benesty, Jingdong Chen, and Yiteng Huang, *Microphone array signal processing*, vol. 1, Springer Science & Business Media, 2008.

[4] Michael E Lockwood and Douglas L Jones, "Beamformer performance with acoustic vector sensors in air," *The Journal of the Acoustical Society of America*, vol. 119, no. 1, pp. 608–619, 2006.

[5] Muawiyath Shujau, CH Ritz, and IS Burnett, "Designing acoustic vector sensors for localisation of sound sources in air," in *Signal Processing Conference, 2009 17th European*. IEEE, 2009, pp. 849–853.

[6] Wenyi Zhang and Bhaskar D Rao, "A two microphone-based approach for source localization of multiple speech sources," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 8, pp. 1913–1928, 2010.

[7] Dovid Levin, Sharon Gannot, and Emanuël AP Habets, "Direction-of-arrival estimation using acoustic vector sensors in the presence of noise," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 105–108.

[8] Dovid Levin, Emanuël AP Habets, and Sharon Gannot, "Maximum likelihood estimation of direction of arrival using an acoustic vector-sensor," *The Journal of the Acoustical Society of America*, vol. 131, no. 2, pp. 1240–1248, 2012.

[9] Dovid Levin, Emanuël AP Habets, and Sharon Gannot, "On the angular error of intensity vector based direction of arrival estimation in reverberant sound fields," *The Journal of the Acoustical Society of America*, vol. 128, no. 4, pp. 1800–1811, 2010.

[10] Malcolm Hawkes and Arye Nehorai, "Wideband source localization using a distributed acoustic vector-sensor array," *Signal Processing, IEEE Transactions on*, vol. 51, no. 6, pp. 1479–1491, 2003.

[11] Dayan Rahamim, Joseph Tabrikian, and Reuven Shavit, "Source localization using vector sensor array in a multipath environment," *Signal Processing, IEEE Transactions on*, vol. 52, no. 11, pp. 3096–3103, 2004.

[12] Huawei Chen and Junwei Zhao, "Coherent signal-subspace processing of acoustic vector sensor array for doa estimation of wideband sources," *Signal Processing*, vol. 85, no. 4, pp. 837–847, 2005.

[13] Chrysostomos L Nikias and Mysore R Raghuveer, "Bispectrum estimation: A digital signal processing framework," *Proceedings of the IEEE*, vol. 75, no. 7, pp. 869–891, 1987.

[14] Philippe Forster and Chrysostomos L Nikias, "Bearing estimation in the bispectrum domain," *Signal Processing, IEEE Transactions on*, vol. 39, no. 9, pp. 1994–2006, 1991.

[15] Zhenghao Shi and Frederick W Fairman, "A comprehensive approach to doa estimation using higher-order statistics," *Circuits, Systems and Signal Processing*, vol. 17, no. 4, pp. 539–557, 1998.

[16] Zhenghao Shi and Frederick W Fairman, "Doa estimation via higher-order cumulants: a generalized approach," in *icassp*. IEEE, 1992, pp. 209–212.

[17] Boaz Porat and Benjamin Friedlander, "Direction finding algorithms based on high-order statistics," *Signal Processing, IEEE Transactions on*, vol. 39, no. 9, pp. 2016–2024, 1991.

[18] Norman Yuen and Benjamin Friedlander, "Doa estimation in multipath: an approach using fourth-order cumulants," *Signal Processing, IEEE Transactions on*, vol. 45, no. 5, pp. 1253–1263, 1997.

[19] Wei Xue, Shan Liang, and Wenju Liu, "Doa estimation of speech source in noisy environments with weighted spatial bispectrum correlation matrix," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 2282–2286.

[20] Chrysostoms L Nikias and Renlong Pan, "Time delay estimation in unknown gaussian spatially correlated noise," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 36, no. 11, pp. 1706–1714, 1988.

[21] Hlrotugu Akaike, "On the use of non-gaussian process in the identification of a linear dynamic system," *Annals of the institute of statistical mathematics*, vol. 18, no. 1, pp. 269–276, 1966.

[22] Hirotugu Akaike, "Note on higher order spectra," *Annals of the Institute of Statistical Mathematics*, vol. 18, no. 1, pp. 123–126, 1966.

[23] M Deistler, "Linear dynamic errors-in-variables models," *Journal of Applied Probability*, pp. 23–39, 1986.

[24] Zdravko I Botev, Joseph F Grotowski, Dirk P Kroese, et al., "Kernel density estimation via diffusion," *The Annals of Statistics*, vol. 38, no. 5, pp. 2916–2957, 2010.

[25] Andrew Varga and Herman JM 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 communication*, vol. 12, no. 3, pp. 247–251, 1993.

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