ROBUST SPEAKER DOA ESTIMATION BASED ON THE INTER-SENSOR DATA RATIO MODEL AND BINARY MASK ESTIMATION IN THE BISPECTRUM DOMAIN

Yanhan Jin[1], Yuexian Zou[1*], C. H. Ritz[2]

[1]ADSPLAB/ELIP, School of ECE, Peking University, Shenzhen, China
[2]SECTE, University of Wollongong, Australia

yanhanjin@pku.edu.cn, zouyx@pkusz.edu.cn, critz@uow.edu.au

ABSTRACT

When noise is directional instead of diffuse, the majority of conventional direction of arrival (DOA) estimation techniques suffer from performance degradation because of mismatched noise models. In this paper, a novel robust DOA estimation algorithm is developed as an initial investigation into DOA estimation of speech under directional non-speech interference (DNSI) and non-directional background noise (NDBN) using an acoustic vector sensor (AVS), a compact co-incident microphone array. Specifically, by defining an inter-sensor data ratio model in the bispectrum domain (BISDR), the relationship between the BISDR and the speech DOA cues are derived. By recursively estimating *a priori* local signal-to-interference ratio of the bispectrum (B-PriLSIR), a robust speech-dominated binary mask (SDBM) is estimated and thus the speech DOA cue is faithfully extracted. Experimental results with simulated and recorded data demonstrate that the proposed algorithm offers high DOA estimation accuracy for all angles and is robust against DNSI and NDBN.

Index Terms— direction of arrival estimation, acoustic vector sensor, directional interference, bispectrum, binary mask estimation

1. INTRODUCTION

The direction of arrival (DOA) estimation in an adverse acoustic environment with surrounding noise has attracted considerable attention due to its wide range of realistic applications such as video conferencing and service robots localizing the speech source swiftly and accurately utilizing compact microphone arrays [1].

The noise may be non-directional (e.g. due to ambient sounds) or directional (e.g. from an interfering source). Existing research shows that the majority of conventional DOA estimation methods [2–6] have been developed under the assumptions that the received noise signals are just non-directional background noise (NDBN). However, in the practical applications, such assumptions may not be realistic and such a model mismatch will degrade the DOA estimation performance.

In existing research investigating DOA estimation in the presence of directional non-speech interference (DNSI), the approach by Nishiura assumes DNSIs to be regular sources, and thus solves the problem in the framework of multi-source DOA estimation [7]. However, this solution requires knowledge of the source number which is often unknown in advance in a real environment. Furthermore, additional effort must be paid to differentiate the speech sources and DNSIs, which is also a difficult task. In [8], a frequency weighting is devised for selecting only speech frequency bands so as to improve the robustness against DNSI. Particularly, derivation of features or cues are required to estimate the DOA. However, for such a frequency domain formulated approach, if the DNSI has a flat frequency distribution, the DOA cues in the frequency domain will be more likely suppressed by the noise, thereby making the DOA estimation less robust. Besides, such an approach often utilizes an array of microphones with a large aperture and presents limits in space-constrained applications.

In our previous work [9], a high resolution speaker DOA estimation algorithm was developed when DNSI and NDBN both exist using a single acoustic vector sensor (AVS). The AVS is an attractive solution for mobile speech applications [10, 11] due to its compact size but ability to accurately record 3D sound without spatial aliasing compared to commonly used scalar sensor arrays [12,13]. By deriving an inter-sensor data ratios model in the bispectrum domain termed as BISDR, we can extract the DOA information of a speaker while at the same time suppress the unwanted DNSI. As an indicator of the higher order statistics (HOS) of a signal, bispectrum of the Gaussian process is always zero [14]. In particular, the bispectrum distribution of speech and non-speech signals are different, which means in the bispectrum domain, most of the speech DOA cues will not be suppressed by DNSI or NDBN [15]. Through analyzing the property of BISDR, the DOA estimation problem is formulated as extracting the reliable speech DOA-related information from the BISDR. Then, a mask is taken to select the speech dominated points in the large-amplitude areas of the bispectra through observation. However, the performance is sensitive to the mask threshold and the kind of interference.

In this paper, we propose a more sophisticated approach to generate the mask to extract the speech DOA cues. A decision-directed method is proposed to derive a speech-dominated binary mask (SDBM) by recursively estimating the *a priori* local signal-to-interference ratio of the bispectrum (B-PriLSIR). Thus, it has a few advantages such as more robustness to the interference. Intensive experiments using various kinds of simulated conditions and recorded data have been carried out to corroborate the effectiveness and robustness of the proposed method.

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

2. DATA MODEL

Generally an AVS is a compact microphone array containing an omni-directional pressure sensor (o-sensor) collocated with three orthogonally oriented pressure gradient sensors (named as u-, v-, w-sensor, respectively). So it is inherently capturing more information than even a few microphones. 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 [5]:

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

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

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

We assume the NDBN is zero-mean additive white Gaussian noise (AWGN) and there is one source of DNSI. Thus the data captured by AVS at time k can be generally expressed as

$$\boldsymbol{x}(k) = \boldsymbol{a}(\theta_s,\phi_s)s(k) * h_s(k) + \boldsymbol{a}(\theta_r,\phi_r)r(k) * h_r(k) + \boldsymbol{n}(k) \quad (3)$$

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 $h_s(k)$, the corresponding manifold vector $\boldsymbol{a}(\theta_s,\phi_s) = [u_s,v_s,w_s,1]^T$; $r(k)$ is the DNSI signal assumed uncorrelated to $s(k)$, with DOA (θ_r,ϕ_r) and the room impulse response $h_r(k)$, the corresponding manifold vector $\boldsymbol{a}(\theta_r,\phi_r) = [u_r,v_r,w_r,1]^T$; $\boldsymbol{n}(k) = [n_u(k),n_v(k),n_w(k),n_o(k)]^T$

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

denotes AWGN at the u-, v-, w-, and o-sensor, respectively.

The data model of the directional l-sensor (where, for simplicity, l refers to u,v,w for compact presentation purpose) and o-sensor in (3) can be respectively further expressed as:

$$x_l(k) = l_s s(k) * h_s(k) + l_r(k) r(k) * h_r(k) + n_l(k) \quad (4)$$

$$x_o(k) = s(k) * h_s(k) + r(k) * h_r(k) + n_o(k) \quad (5)$$

3. PROPOSED DOA ESTIMATION METHOD

3.1. Bispectrum Domain Representation

In this subsection, we will derive the AVS data model in bispectrum domain.

As the $n_l(k)$ and $n_o(k)$ are zero-mean Gaussian, their bispectra are identical to zero. Under the assumption that s, r and n are uncorrelated with each other, and $s_h(k)$=$s(k)$*$h_s(k)$, $r_h(k)$=$r(k)$*$h_r(k)$, according to the derivation in [16], the cross-bispectrum between $x_l(k)$ and $x_o(k)$ and the bispectrum of $x_o(k)$ can be expressed as:

$$B_{x_o x_l x_o}(\Omega_1,\Omega_2) = l_s B_{s_h s_h s_h}(\Omega_1,\Omega_2) + l_r B_{r_h r_h r_h}(\Omega_1,\Omega_2) \quad (6)$$

$$B_{x_o x_o x_o}(\Omega_1,\Omega_2) = B_{s_h s_h s_h}(\Omega_1,\Omega_2) + B_{r_h r_h r_h}(\Omega_1,\Omega_2) \quad (7)$$

where $B_{s_h s_h s_h}(\Omega_1,\Omega_2)$ is the bispectrum of $s_h(k)$ with the corresponding manifold vector component l_s (where l_s refers to u_s, v_s, w_s, respectively). Similarly, $B_{r_h r_h r_h}(\Omega_1,\Omega_2)$ is the bispectrum of $r_h(k)$ with the corresponding manifold vector component l_r (where l_r refers to u_r, v_r, w_r, respectively).

Note that the first terms in (6) and (7) are only related to the speech source, and the second terms are only related to the DNSI.

3.2. Bispectrum Inter-Sensor Data Ratio (BISDR)

Following the derivation of the sensor data ratio idea of [5], in this subsection, we define the BISDR of the AVS in the bispectrum domain as follows:

$$I_{lo}(\Omega_1,\Omega_2) \triangleq B_{x_o x_l x_o}(\Omega_1,\Omega_2) / B_{x_o x_o x_o}(\Omega_1,\Omega_2), l = u,v,w \quad (8)$$

where $I_{lo}(\Omega_1,\Omega_2)$ is termed as the BISDR between l- and o-sensor. Substituting (6) and (7) into (8) gives

$$\begin{aligned} I_{lo}(\Omega_1,\Omega_2) &= \frac{l_s B_{s_h s_h s_h}(\Omega_1,\Omega_2) + l_r B_{r_h r_h r_h}(\Omega_1,\Omega_2)}{B_{s_h s_h s_h}(\Omega_1,\Omega_2) + B_{r_h r_h r_h}(\Omega_1,\Omega_2)} \\ &= \frac{l_s [B_{s_h s_h s_h}(\Omega_1,\Omega_2) + B_{r_h r_h r_h}(\Omega_1,\Omega_2)]}{B_{s_h s_h s_h}(\Omega_1,\Omega_2) + B_{r_h r_h r_h}(\Omega_1,\Omega_2)} \\ &+ \frac{l_r B_{r_h r_h r_h}(\Omega_1,\Omega_2) - l_s B_{r_h r_h r_h}(\Omega_1,\Omega_2)}{B_{s_h s_h s_h}(\Omega_1,\Omega_2) + B_{r_h r_h r_h}(\Omega_1,\Omega_2)} \end{aligned} \quad (9)$$

To simplify (9), rewrite it as follows:

$$I_{lo}(\Omega_1,\Omega_2) = l_s + \varepsilon_l(\Omega_1,\Omega_2) \quad (10)$$

the second term in (10) is the residual term given by

$$\varepsilon_l(\Omega_1,\Omega_2) = \frac{l_r - l_s}{1 + B_{s_h s_h s_h}(\Omega_1,\Omega_2) / B_{r_h r_h r_h}(\Omega_1,\Omega_2)} \quad (11)$$

From (10), the compact expression of BISDR can be written as:

where $$\boldsymbol{I}(\Omega_1,\Omega_2) = \boldsymbol{b}(\theta_s,\phi_s) + \boldsymbol{\varepsilon}(\Omega_1,\Omega_2) \quad (12)$$

$$\boldsymbol{I}(\Omega_1,\Omega_2) = [I_{uo}(\Omega_1,\Omega_2), I_{vo}(\Omega_1,\Omega_2), I_{wo}(\Omega_1,\Omega_2)]^T \quad (13)$$

$$\boldsymbol{b}(\theta_s,\phi_s) = [u_s, v_s, w_s]^T \quad (14)$$

$$\boldsymbol{\varepsilon}(\Omega_1,\Omega_2) = [\varepsilon_u(\Omega_1,\Omega_2), \varepsilon_v(\Omega_1,\Omega_2), \varepsilon_w(\Omega_1,\Omega_2)]^T \quad (15)$$

Obviously, the first term $\boldsymbol{b}(\theta_s,\phi_s)$ in (12) is just the compact expression of the speech direction cosines in the manifold vector $\boldsymbol{a}(\theta_s,\phi_s)$, and we refer to "speech DOA cue".

Ideally, to obtain the speech DOA cue from $\boldsymbol{I}(\Omega_1,\Omega_2)$ in (12), the unwanted residual term $\boldsymbol{\varepsilon}(\Omega_1,\Omega_2)$ is supposed to be zero. Analyzing (11), note that if we find the $(\Omega_{1g},\Omega_{2g})$ (termed as frequency points, FPs) where $B_{s_h s_h s_h}(\Omega_{1g},\Omega_{2g}) / B_{r_h r_h r_h}(\Omega_{1g},\Omega_{2g}) \gg 1$, then $\boldsymbol{\varepsilon}(\Omega_{1g},\Omega_{2g})$ approaches zero. Besides, it is obvious that the speech DOA cue is possible at the FPs where the value of $B_{s_h s_h s_h}(\Omega_{1g},\Omega_{2g})$ is large. Accordingly, these speech FPs with high local speach-to-interference ratio can be termed as HLSIR-SFPs, where the speech DOA cue can be faithfully obtained. With the discussion above, the remaining task is to determine these HLSIR-SFPs properly.

3.3. Speech Dominated Binary Mask (SDBM) Estimation

The binary masking technique is one of the efficient techniques for extracting the HLSIR-SFPs, where a metric is used to judge whether or not the FP is speech dominated. In this subsection, we will propose a dual strategy to estimate a SDBM in bispectrum domain.

Step 1: Estimation of HLSIR mask.

Our idea is triggered by the decision-directed *a priori* SNR estimator proposed by Ephraim and Malah for speech amplitude estimator [17] where a ratio between the short-time spectral component of speech and noise is defined and updated on the basis of a previous amplitude estimate.

Considering that the HLSIR-SFPs in this study are analogous to the FPs with high values of the ratio between the amplitude bispectrum component of speech and DNSI, we make an effort to estimate the SDBM with a similar *a priori* SIR estimator.

Since we work with the AVS, there are four channel signals that can be manipulated to design the mask. From (6) and (7), the speech bispectrum components in $B_{x_o x_l x_o}(\Omega_1,\Omega_2)$ and $B_{x_o x_o x_o}(\Omega_1,\Omega_2)$ differ only up to a scale factor l_s. Therefore, these speech components have the same distributions in the amplitude bispectrum. Similarly, so do the DNSI components. As a result, if an amplitude bispectrum FP for $o-$sensor of AVS $B_{x_o x_o x_o}(\Omega_1,\Omega_2)$ is HLSIR-SFP, the corresponding FP for the other three sensors $B_{x_o x_l x_o}(\Omega_1,\Omega_2)$ is also HLSIR-SFP, and vice versa. In this study, we therefore take $B_{x_o x_o x_o}(\Omega_1,\Omega_2)$ for the mask estimation. Actually, informal experiments omitted for brevity also shows that $B_{x_o x_l x_o}(\Omega_1,\Omega_2)$ gives almost the same result.

In order to calculate the ratio between the amplitude bispectrum component of speech and DNSI, we define the *a priori* local signal-to-interference ratio of the bispectrum (B-PriLSIR) $\xi(\Omega_1,\Omega_2)$ and *a posteriori* local signal-to-interference ratio of the bispectrum (B-PostLSIR) $\gamma(\Omega_1,\Omega_2)$ as follows, respectively:

$$\xi(\Omega_1,\Omega_2) \triangleq \frac{|B_{s_h s_h s_h}(\Omega_1,\Omega_2)|^2}{\lambda_r(\Omega_1,\Omega_2)} \quad (16)$$

$$\gamma(\Omega_1,\Omega_2) \triangleq \frac{|B_{x_o x_o x_o}(\Omega_1,\Omega_2)|^2}{\lambda_r(\Omega_1,\Omega_2)} \quad (17)$$

where $\lambda_r(\Omega_1,\Omega_2) \triangleq E\left\{|B_{r_h r_h r_h}(\Omega_1,\Omega_2)|^2\right\}$ is the estimation of the bispectrum power of DNSI. It can be obtained from nonspeech intervals (assuming the speech is not active in the initial 0.3s for faciliation.). Actually, the nonspeech intervals can also be judged utilizing voice activity detection algorithm (VAD) [18]. $B_{s_h s_h s_h}(\Omega_1,\Omega_2)$ and $B_{x_o x_o x_o}(\Omega_1,\Omega_2)$ are respectively the bispectrum power of the speech signal received by the AVS before and after it is polluted by the DNSI. Assuming the speech and DNSI are uncorrelated, we have

$$\xi(\Omega_1,\Omega_2) = \gamma(\Omega_1,\Omega_2) - 1 \quad (18)$$

Following (16) and (18), we note that $B_{s_h s_h s_h}(\Omega_1,\Omega_2)$ is unknown. Instead, $B_{x_o x_o x_o}(\Omega_1,\Omega_2)$ can be directly computed from the current $(t)^{th}$ segment of AVS received signal x_o. Hence, we estimate $\xi^{(t)}(\Omega_1,\Omega_2)$ taking use of its relation to $\gamma^{(t)}(\Omega_1,\Omega_2)$ in (18), where the superscript $(\cdot)^{(t)}$ indicates the time segment t and each data segment is taken from several consecutive time frames, whose details will be explained in Section IV. Inspired by the "decision-directe" method in [17], we consider recursively updating $\xi^{(t)}(\Omega_1,\Omega_2)$ using the amplitude estimator of the $(t-1)^{th}$ segment for smoothing to reduce speech distortion [17] and the B-PostLSIR of $(t)^{th}$ segment to reduce the residual noise. Then the B-PriLSIR $\xi(\Omega_1,\Omega_2)$ for the current $(t)^{th}$ segment can be estimated as

$$\hat{\xi}^{(t)}(\Omega_1,\Omega_2) = \beta \frac{\left|B^{(t-1)}_{s_h s_h s_h}(\Omega_1,\Omega_2)\right|^2}{\lambda_r(\Omega_1,\Omega_2)} + (1-\beta) P\left[\gamma^{(t)}(\Omega_1,\Omega_2) - 1\right] \quad (19)$$

where $P[\cdot]$ denotes half-wave rectification to ensure the estimated B-PriLSIR to be positive. $\beta \in [0,1]$ is a forgetting factor, which is set to be 0.7 by empirical results. Followling (16), the $\left|B^{(t-1)}_{s_h s_h s_h}(\Omega_1,\Omega_2)\right|^2$ in (19) can be approximated as:

$$\begin{aligned}\left|\hat{B}^{(t-1)}_{s_h s_h s_h}\right|^2 &= \left(\left|\hat{B}^{(t-1)}_{s_h s_h s_h}\right|^2 / \left|\hat{B}^{(t-1)}_{x_o x_o x_o}\right|^2\right)\cdot\left|\hat{B}^{(t-1)}_{x_o x_o x_o}\right|^2 \\ &= \left[\left|\hat{B}^{(t-1)}_{s_h s_h s_h}\right|^2 / \left(\left|\hat{B}^{(t-1)}_{s_h s_h s_h}\right|^2 + \lambda_r\right)\right]\cdot\left|\hat{B}^{(t-1)}_{x_o x_o x_o}\right|^2 \\ &= \left[\hat{\xi}^{(t-1)}/(\hat{\xi}^{(t-1)}+1)\right]\cdot\left|\hat{B}^{(t-1)}_{x_o x_o x_o}\right|^2\end{aligned} \quad (20)$$

Substituting (20) into (19), we get

$$\hat{\xi}^{(t)} = \beta\hat{\gamma}^{(t-1)}\frac{\hat{\xi}^{(t-1)}}{\hat{\xi}^{(t-1)}+1} + (1-\beta)P\left[\hat{\gamma}^{(t)}-1\right] \quad (21)$$

In (20) and (21), "(Ω_1,Ω_2)" is omitted for simplicity.

From the definition of $\xi(\Omega_1,\Omega_2)$ in (16), it is easy to see that $\xi(\Omega_{1g},\Omega_{2g}) \gg 1$ implies $B_{s_h s_h s_h s_h}(\Omega_{1g},\Omega_{2g}) / B_{r_h r_h r_h r_h}(\Omega_{1g},\Omega_{2g}) \gg 1$ with high probability. Then, the FPs with HLSIR can be selected as the ones with large $\hat{\xi}^{(t)}(\Omega_1,\Omega_2)$:

$$m_\xi(\Omega_1,\Omega_2) \triangleq \begin{cases} 1 & \hat{\xi}^{(t)}(\Omega_1,\Omega_2) > \zeta \\ 0 & \text{otherwise} \end{cases} \quad (22)$$

where ζ is a threshold. The larger ζ is , the less polluted FPs are selected, but the amount of selected FPs gets fewer. We empirically set it as 1.5 considering different DNSI conditions.

However, evaluating $\xi(\Omega_1,\Omega_2)$ in (16), we noted that there are two cases to get $\xi(\Omega_1,\Omega_2) \gg 1$:

Case 1: $B_{s_h s_h s_h s_h}(\Omega_{1g},\Omega_{2g})$ is large, $B_{r_h r_h r_h r_h}(\Omega_{1g},\Omega_{2g})$ is small;

Case 2: $B_{s_h s_h s_h s_h}(\Omega_{1g},\Omega_{2g})$ is small, $B_{r_h r_h r_h r_h}(\Omega_{1g},\Omega_{2g})$ is smaller.

Obviously, large value of $\xi(\Omega_1,\Omega_2)$ is not a perfect indicator of speech dominated FPs since only *Case 1* is preferred.

As a result, it is straightforward that we should select the speech frequency points (SFPs) where the speech bispectrum power $|B_{s_h s_h s_h}(\Omega_1,\Omega_2)|^2$ is of large value conforming to *Case 1* and we can determine the HLSIR-SFPs accordingly.

Step 2: Estimation of SDBM.

Following (16), with the estimated B-PriLSIR $\hat{\xi}^{(t)}(\Omega_1,\Omega_2)$, we can directly estimate $|B_{s_h s_h s_h}(\Omega_1,\Omega_2)|^2$ at current segment as:

$$\left|\hat{B}^{(t)}_{s_h s_h s_h}(\Omega_1,\Omega_2)\right|^2 = \hat{\xi}^{(t)}(\Omega_1,\Omega_2)\cdot\lambda_r(\Omega_1,\Omega_2) \quad (23)$$

Thereupon the SFPs can be selected with binary mask $m_s(\Omega_1,\Omega_2)$:

$$m_s(\Omega_1,\Omega_2) \triangleq \begin{cases} 1 & \left|\hat{B}^{(t)}_{s_h s_h s_h}(\Omega_1,\Omega_2)\right|^2 > M \max\limits_{(\Omega_1\Omega_2)}\left(\left|\hat{B}^{(t)}_{s_h s_h s_h}(\Omega_1,\Omega_2)\right|^2\right) \\ 0 & \text{otherwise} \end{cases} \quad (24)$$

where a threshold $M > 0$ is applied to indicate the presence of SFP. While too large value can cause the insufficiency of selected FPs. It is set to be 0.15 by empirical results with the details in Section IV.

With the estimated $m_\xi(\Omega_1,\Omega_2)$ and $m_s(\Omega_1,\Omega_2)$, we further estimate the SDBM satisfying the statement of *Case 1* as:

$$m(\Omega_1,\Omega_2) \triangleq m_\xi(\Omega_1,\Omega_2)\cdot m_s(\Omega_1,\Omega_2) \quad (25)$$

Logically, the HLSIR-SFPs correspond to the non-zero values of $m(\Omega_1,\Omega_2)$ estimated in (25). Then the BISDR at these HLSIR-SFPs is given by:

$$\widetilde{\boldsymbol{I}}(\Omega_1,\Omega_2) = m(\Omega_1,\Omega_2)\cdot\boldsymbol{I}(\Omega_1,\Omega_2) \quad (26)$$

where $\widetilde{\boldsymbol{I}}(\Omega_1,\Omega_2)$ is the masked BISDR. With the derivation of (26) and (25), from (12), we can reach the following approximation:

$$\widetilde{\boldsymbol{I}}(\Omega_1,\Omega_2) = m(\Omega_1,\Omega_2)\cdot\boldsymbol{b}(\theta_s,\phi_s) + m(\Omega_1,\Omega_2)\cdot\boldsymbol{\varepsilon}(\Omega_1,\Omega_2) \approx \boldsymbol{b}(\theta_s,\phi_s) \quad (27)$$

As shown in (27), the speech DOA cue $\boldsymbol{b}(\theta_s,\phi_s)$ can be extracted by the estimated mask $m(\Omega_1,\Omega_2)$.

To validate our derivation, we visualize the bispectrum of speech and DNSI, the $I_{uo}(\Omega_1,\Omega_2)$ and $\widetilde{I}_{uo}(\Omega_1,\Omega_2)$ in log-scale in Fig. 1. Comparing Fig. 1(a) and (b), speech and DNSI have different bispectrum patterns with some overlapping areas. Besides, the spatial pattern shown in Fig. 1(d) is similar to the high energy one in Fig. 1(a), which indicates significant speech information is extracted from Fig. 1(c). Moreover, $\widetilde{I}_{uo}(\Omega_1,\Omega_2)$ in Fig. 1(d) has similar values at the red dots, which validates the result derived in (27).

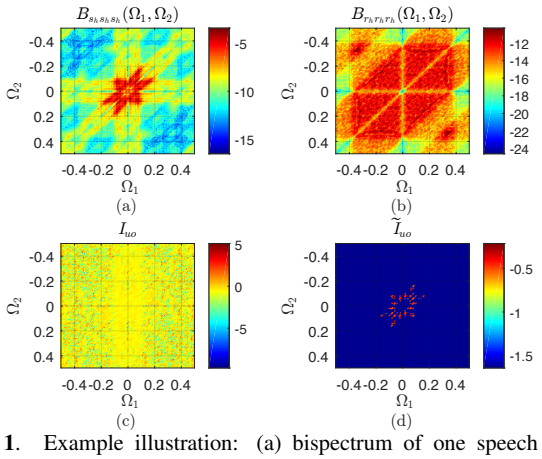

Fig. 1. Example illustration: (a) bispectrum of one speech signal, (b) bispectrum of one DNSI signal, (c) $I_{uo}(\Omega_1,\Omega_2)$ by (8), (d) $\widetilde{I}_{uo}(\Omega_1,\Omega_2)$ by (26). (DNSI is hfchannel noise, SIR = 10dB; AWGN, SNR = 10dB.)

3.4. The Proposed DOA Estimation Algorithm

From the description above, it is obvious that the masked BISDR $\widetilde{\boldsymbol{I}}(\Omega_1,\Omega_2)$ can be viewed as random variables in bispectrum domain with mean of u_x, v_s and w_s, respectively. Specifically, the DOA estimation task is to estimate the cluster center at (u_s,v_s,w_s) by clustering the masked BISDR $\widetilde{\boldsymbol{I}}(\Omega_1,\Omega_2)$ corresponding to all HLSIR-SFPs. To achieve an effective and robust clustering result, the kernel density estimation (KDE) method of [19] is adopted. 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

$$\hat{\theta}_s = \cos^{-1}\hat{w}_s,\ \hat{\phi}_s = \tan^{-1}(\hat{v}_s/\hat{u}_s) \quad (28)$$

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

1) Segment the AVS output data $\boldsymbol{x}(k)$ and calculate the bispectrum of the four sensors $B_{x_o x_l x_o}(\Omega_1,\Omega_2)$ and $B_{x_o x_o x_o}(\Omega_1,\Omega_2)$ in each time segment by (6) and (7).

2) Calculate the BISDR $I_{lo}(\Omega_1,\Omega_2)$ between sensors by (8).

3) Get the SDBM $m(\Omega_1,\Omega_2)$ by (25) and then the masked BISDR $\widetilde{\boldsymbol{I}}(\Omega_1,\Omega_2)$ by (26).

4) Estimate the DOA $(\hat{\theta}_s,\hat{\phi}_s)$ via (28) by the clustering result $(\hat{u}_s,\hat{v}_s,\hat{w}_s)$ derived using KDE [19].

4. EXPERIMENTAL RESULTS

In this section, several experiments are carried out to evaluate the performance of our proposed AVS-MBISDR algorithm under different conditions. Another three methods capable of DOA estimation using a small microphone array are taken as the comparison methods, including the well-known GMDA-Laplace algorithm [12] and our previously proposed AVS-ISDR algorithm [5] and AVS-BISDR [9].

Throughout the simulations, the speech signal is of 3 seconds and sampled at $8kHz$ from TIMIT [20]. One DNSI is set at $(60°,75°)$. Unless otherwise specified, the speech source is set at $(60°,45°)$ with no reverberation. The DNSI is taken from Noisex92 [21]. In addition, the AWGN is taken as NDBN with SNR = 10dB. For processing signals, the frame size is set to be 256 samples with 60% overlap. For the GMDA-Laplace algorithm, following the setup in [12], the DOA

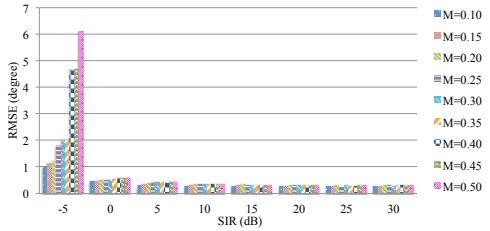

Fig. 2. RMSE versus SIR levels using different mask thresholds. (DNSI is factory noise)

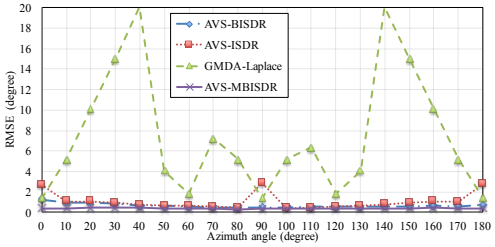

Fig. 3. RMSE versus different source azimuth angles. (DNSI is machinegun noise, SIR = 5dB)

estimation results are obtained by running the algorithm twice since originally it just uses two microphones.

The performance metric used is the root mean squared error (RMSE) of the speech source, averaging over 100 independent trials with different random AWGN. It is defined as $RMSE = 0.5\sqrt{\sum_{j=1}^{100}((\hat{\theta}_j-\theta)^2+(\hat{\phi}_j-\phi)^2)/100}$ where $\hat{\theta}_j$ and $\hat{\phi}_j$ are respectively the estimated angles of the speaker θ and ϕ on the j^{th} trial.

1) Effect of the mask threshold: This experiment aims to evaluate the impact of choosing different threshold M for SDBM on the performance of AVS-MBISDR under different SIR conditions since M is an important parameter. The results are shown in Fig. 2. It is noted that when SIR<0dB and M>0.35, the RMSE is large. This is reasonable because the larger M is, the fewer the number of selected FPs. In low SIR conditions, with the strong effect of DNSI, enough FPs are necessary for the estimation performance. While when SIR>0dB, the RMSE is not sensitive to M. The optimal M can be selected as 0.15, which gives the best results under most SIR and DNSI conditions.

2) RMSE versus azimuth angle: This experiment is conducted to evaluate the sensitivity of the proposed DOA algorithm over different azimuth angles. We fix $\theta_s = 60^\circ$. The experimental results are shown in Fig. 3. We are encouraged to see that our proposed AVS-MBISDR algorithm outperforms the comparison algorithms, where the RMSE values are constantly closed to 0° for all angles. It is noted that the estimated DOA by the comparison algorithm gives larger fluctuation from its true DOA when the azimuth goes closer to some special angles (e.g. 0° and 180°).

3) RMSE versus different SIR and DNSI: In this experiment, the behavior of AVS-MBISDR under different SIR and types of DNSI is evaluated. Experimental results are presented in Fig. 4. As expected, the RMSE of the proposed method keeps as a small constant (close to 0°) even when the SIR is less than 0dB. And also it works better than the previous work AVS-BISDR for certain types of noises which are less stationary. The other two algorithms, by contrast, both suffer a severe decline of the DOA estimation performance with the impact of strong DNSI, as it is hard to differentiate the speech source and DNSI in time-frequency spectrum. This verifies our proposed algorithm is more effective and robust against DNSI.

4) RMSEs versus different reverberation levels: This experiment aims at evaluating the influence of the reverberation on the performance of the proposed AVS-MBISDR. The room impulse response is simulated by the image method [22] with the virtual rectangular room size of $10\times5\times4m^3$, and five different reverberation time (RT_{60}) conditions are considered. It is seen in Fig. 5 that the proposed algorithm

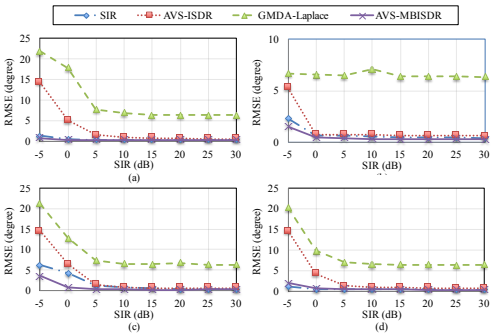

Fig. 4. RMSE versus different SIR levels and DNSI signals as: white Gaussian noise (a), machinegun noise (b), F16 noise (c), and factory noise (d).

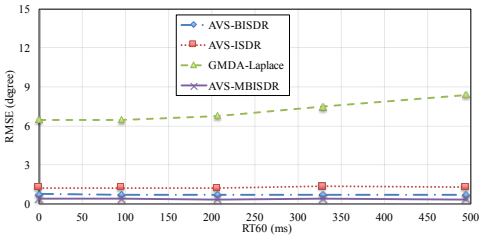

Fig. 5. RMSE versus different RT_{60}. (DNSI is machinegun noise, SIR = 5dB)

Table 1: DOA Estimation Resluts in a Real Scenario

True DOA	θ	90°	90°	90°	90°	90°
	ϕ	0°	45°	90°	135°	180°
AVS-MBISDR	θ	90.66°	89.71°	90.34°	90.72°	89.78°
	ϕ	1.92°	43.82°	90.95°	139.94°	179.93°

has superior estimation accuracy over AVS-BISDR and AVS-ISDR for all RT_{60} conditions under this setup. This indicates that our proposed algorithm is not sensitive to room reverberation, which is a favorable property since the performance of many existing algorithms, such as GMDA-Laplace, degrades when heavy room reverberation exists.

5) DOA estimation based on recorded data: This experiment is conducted to evaluate the performance of the proposed AVS-MBISDR algorithm in a real scenario with realtime noise, interferences and reverberation using the AVS data capturing system developed by AD-SPLAB [5]. The environment is as follows: The room is about $8.5\times3\times5m^3$ with the uncontrolled background noise including air conditioning and computer servers that can be viewed as the DNSI. The SNR measured is approximately 20dB and the reverberation is present. The distance between the speaker and the AVS is $0.5m$. Five different groups of DOA are estimated respectively in Table 1, one recorded sentence from male and one from female in each group. We can see that the DOA estimation errors of the proposed AVS-MBISDR algorithm are less than 5° in each group with real recorded data.

5. CONCLUSIONS

In this paper, using a single AVS, a novel robust 3-D DOA estimation method (termed as AVS-MBISDR) is developed in bispectrum domain for speech sources. To guarantee robust speech HOS spatial location information extraction against DNSI, the SDBM is further obtained. Theoretical analysis and experimental results with the simulated and real captured data illustrate that AVS-MBISDR exhibits excellent estimation performance under various DNSI conditions even when SIR is smaller than 0dB. Also, superior performance is achieved when reverberation is strong. Further research is required to estimate the effectiveness of the approach for multiple source DOA estimation and the ability to operate in real-time on embedded hardware.

6. REFERENCES

[1] 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.

[2] Michael Brandstein and Darren Ward, *Microphone arrays: signal processing techniques and applications*, Springer Science & Business Media, 2001.

[3] Jacek P Dmochowski, Jacob Benesty, and Sofiene Affes, "A generalized steered response power method for computationally viable source localization," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 8, pp. 2510–2526, 2007.

[4] Mengqi Ren and Yue Xian Zou, "A novel multiple sparse source localization using triangular pyramid microphone array," *Signal Processing Letters, IEEE*, vol. 19, no. 2, pp. 83–86, 2012.

[5] 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.

[6] Maximo Cobos, Jose J Lopez, and Sascha Spors, "A sparsity-based approach to 3d binaural sound synthesis using time-frequency array processing," *EURASIP Journal on Advances in Signal Processing*, vol. 2010, pp. 2, 2010.

[7] Takanobu Nishiura, Satoshi Nakamura, and Kiyohiro Shikano, "Talker localization in a real acoustic environment based on doa estimation and statistical sound source identification," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*. IEEE, 2002, vol. 1, pp. I–893.

[8] Wei Xue, Shan Liang, and Wenju Liu, "Interference robust doa estimation of human speech by exploiting historical information and temporal correlation.," in *INTERSPEECH*, 2013, pp. 2895–2899.

[9] Y H Jin and YX Zou, "Robust speaker doa estimation with single avs in bispectrum domain," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 3196–3200.

[10] 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.

[11] 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.

[12] 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.

[13] Noam R Shabtai, Boaz Rafaely, and Yaniv Zigel, "The effect of reverberation on optimal gmm order and cms performance in speaker verification systems," *SCIYO. COM*, p. 37, 2010.

[14] David R Brillinger, "An introduction to polyspectra," *The Annals of mathematical statistics*, pp. 1351–1374, 1965.

[15] 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.

[16] 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.

[17] Yariv Ephraim and David Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 32, no. 6, pp. 1109–1121, 1984.

[18] Jongseo Sohn, Nam Soo Kim, and Wonyong Sung, "A statistical model-based voice activity detection," *Signal Processing Letters, IEEE*, vol. 6, no. 1, pp. 1–3, 1999.

[19] 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.

[20] John S Garofolo, Linguistic Data Consortium, et al., *TIMIT: acoustic-phonetic continuous speech corpus*, Linguistic Data Consortium, 1993.

[21] 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.

[22] 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.