MULTISOURCE DOA ESTIMATION BASED ON TIME-FREQUENCY SPARSITY AND JOINT INTER-SENSOR DATA RATIO WITH SINGLE ACOUSTIC VECTOR SENSOR

Y. X. ZOU[1], *Wei SHI*[1], *Bo Li*[1], *C. H. Ritz*[2], *M. Shujau*[2] and *Jiangtao XI*[2]
[1]ADSP LAB School of Electronic Computer Engineering, Peking University, Shenzhen 518055, China
[2]School of Electrical, Computer, and Telecommunications Engineering, University of Wollongong, Australia
{zouyx@pkusz.edu.cn, shiwei@sz.pku.edu.cn, critz@uow.edu.au and jiangtao@uow.edu.au}

ABSTRACT

By exploring the time-frequency (TF) sparsity property of the speech, the inter-sensor data ratios (ISDRs) of single acoustic vector sensor (AVS) have been derived and investigated. Under noiseless condition, ISDRs have favorable properties, such as being independent of frequency, DOA related with single valuedness, and no constraints on near or far field conditions. With these observations, we further investigated the behavior of ISDRs under noisy conditions and proposed a so-called ISDR-DOA estimation algorithm, where high local SNR data extraction and bivariate kernel density estimation techniques have been adopted to cluster the ISDRs representing the DOA information. Compared with the traditional DOA estimation methods with a small microphone array, the proposed algorithm has the merits of smaller size, no spatial aliasing and less computational cost. Simulation studies show that the proposed method with a single AVS can estimate up to seven sources simultaneously with high accuracy when the SNR is larger than 15dB. In addition, the DOA estimation results based on recorded data further validates the proposed algorithm.

Index Terms—Direction-of-arrival estimation, acoustic vector sensor, sinusoidal modeling, time-frequency sparsity, kernel density estimation

1. INTRODUCTION

Direction of arrival (DOA) estimation for multiple spatial speech sources using a small microphone array (SMA) has wide applications in many emerging areas, such as hands-free mobile communication, due to its compact size. Research has shown that traditional DOA estimation algorithms generally lack the ability in providing high DOA estimation accuracy when sources are located close to the end-fire. Moreover, they are not able to estimate the DOAs when the number of sources is larger than the number of microphones used (which is known as the underdetermined problem) [1]-[3].

Recently, several DOA methods with SMA have been proposed to solve the underdetermined problem. In principle, they utilize the time-frequency (TF) domain sparsity of the sources where the Inter-sensor Phase Difference (IPD) or Inter-sensor Time Difference (ITD) can be estimated in each TF point [4]. Correspondingly, several multiple source DOA estimation algorithms have been developed by using different techniques to explore the IPD/ITD information obtained [1-6]. Besides, a TF-sparsity based 3D binaural sound localization using IPD was also recently presented [7]. Motivated by the multiple source DOA estimation capability and the high estimation accuracy of the IPD/ITD-based DOA estimation methods for sources located close to the broadside, we also proposed a new solution with a 3D SMA called a triangular pyramid microphone array-TPMA to achieve better DOA estimation performance for all angles [8]. However, our solution has an obvious downside compared with two microphone solutions since we need four omnidirectional microphones and this limits it applications when the geometry size is the main concern. To combat this, we turn to the acoustic vector sensor (AVS) technique, which has a long research history and numerous achievements in underwater acoustic source localization applications [9]. The distinctive properties and advantages of the AVS over SMA come from its special sensor co-locating structure. One AVS consists of one omnidirectional sensor with two to three orthogonally oriented directional sensors, where particle velocity sensors or differential microphones are often used as the directional sensors. Hence, by exploring the inter-relation of the signals received by different AVS sensors, DOA estimation methods designed for traditional microphone array have been extended to the AVS or AVS array. For example, an ESPRIT-based DOA estimation method with a single vector hydrophone was proposed in [10], which could resolve up to four narrow-band sources. Recently, the AVS technique also gains great attention in speech enhancement applications [13].

Following the spirit of time-frequency sparsity based DOA estimation methods with SMA and the favorable properties of the AVS, we make an effort to develop a more practical DOA estimation technique with a single AVS for real environment DOA applications. The novelty of our solution lies in the investigation of the inter-sensor data ratio (ISDR) within a single AVS together with the high local SNR TF point extraction and the bivariate kernel density estimation (KDE) techniques. Initial performance analysis showed that our proposed multiple sources DOA estimation method with a single AVS brings three major merits: 1) The proposed approach provides the ability to estimate the elevation and azimuth at the same time with a single AVS, which is of much smaller physical size than the traditional SMA; 2) Because the ISDRs of an AVS are independent to the source frequencies, there will be no need to consider the spatial aliasing problem discussed in [3]; 3) The underlying

This work is partially supported by National Natural Science Foundation of China (No: 61271309) and the Shenzhen Science & Technology Fundamental Research Program (No: JCY201110006).

clusters are simplified to points instead of lines, and thus clustering methods such as the histogram and K-means techniques can be employed, which requests less computational cost than the GMDA algorithm [3]. Experiments showed the good performance and the effectiveness of our proposed multisource DOA estimation method.

2. MULTISOURCE DOA ESTIMATION WITH SINGLE AVS

In this section, the AVS data model will be presented first. Our proposed multisource DOA estimation with single AVS will be derived based on TF domain sparsity accordingly.

2.1. AVS Data Mode

Supposing there are *K* acoustic signals $s_k(t)$ (k=1,…,K) impinging upon an AVS composed of *four* sensors, where the omnidirectional sensor and three directional sensors are defined as the *o*-, *u*-, *v*- and *w*-sensor, respectively. A *four*-sensor AVS has the following 4×1 manifold vector for the *k*-th spatial source with DOA of $\{\theta_k, \phi_k\}$ [12]:

$$\mathbf{a}(\theta_k, \phi_k) \equiv \begin{bmatrix} 1 & u_k & v_k & w_k \end{bmatrix}^T \quad (1)$$

where $\theta_k \in [0°, 180°)$ and $\phi_k \in [0°, 360°)$ are respectively the elevation and azimuth angles. u_k, v_k and w_k are named as *x*-, *y*- and *z*-axis direction cosines respectively, given by:

$$u_k = \sin\theta_k \cos\phi_k, v_k = \sin\theta_k \sin\phi_k, w_k = \cos\theta_k \quad (2)$$

It is noted that u_k, v_k and w_k are only dependent on the DOA of the *k*-th source and are independent of the signal frequencies. Assuming there are *K* spatial sources, the output of an AVS at time *t* can be modeled as:

$$x_o(t) = \sum_{k=1}^{K} s_k(t) + n_o(t) \quad (3)$$

$$x_u(t) = \sum_{k=1}^{K} u_k s_k(t) + n_u(t) \quad (4)$$

$$x_v(t) = \sum_{k=1}^{K} v_k s_k(t) + n_v(t) \quad (5)$$

$$x_w(t) = \sum_{k=1}^{K} w_k s_k(t) + n_w(t) \quad (6)$$

where $x_o(t)$, $x_u(t)$, $x_v(t)$ and $x_w(t)$ denote the output of the *o*-, *u*-, *v*-, *w*-sensor; $n_o(t)$, $n_u(t)$, $n_v(t)$ and $n_w(t)$ are the additive zero-mean Gaussian noise at the *o*-, *u*-, *v*-, *w*-sensor respectively, which are assumed uncorrelated to each other and uncorrelated to speech sources.

2.2 The Proposed Inter-Sensor Data Ratio (ISDR)

The major foundation of our proposed multisource DOA estimation method lays on the time-frequency (TF) domain sparsity in speech signals, which is commonly and widely accepted in practical applications [3]. Therefore, it is most probably true that, at a specific TF point (τ, ω), only one speech source with the highest energy dominates and the contributions from other sources can be negligible. Hence, let's suppose that the *k*-th source is dominant at the TF point (τ, ω). Thus, taking the short-time Fourier transform (STFT) of (3)-(6) gives

$$X_o(\tau, \omega) = S_k(\tau, \omega) + N_o(\tau, \omega) \quad (7)$$

$$X_u(\tau, \omega) = u_k S_k(\tau, \omega) + N_u(\tau, \omega) \quad (8)$$

$$X_v(\tau, \omega) = v_k S_k(\tau, \omega) + N_v(\tau, \omega) \quad (9)$$

$$X_w(\tau, \omega) = w_k S_k(\tau, \omega) + N_w(\tau, \omega) \quad (10)$$

The inter-sensor data ratio (ISDR) between the *u*- and the *o*-sensor can be defined as follows

$$I_{uo}(\tau, \omega) \triangleq X_u(\tau, \omega) \big/ X_o(\tau, \omega) \quad (11)$$

Similarly, we have the following ISDRs defined between *v*-, *w*- and *o*-sensor:

$$I_{vo}(\tau, \omega) = X_v(\tau, \omega) \big/ X_o(\tau, \omega) \quad (12)$$

$$I_{wo}(\tau, \omega) = X_w(\tau, \omega) \big/ X_o(\tau, \omega) \quad (13)$$

Substituting (8) and (7) into (11) gives

$$I_{uo}(\tau, \omega) = \frac{u_k S_k(\tau,\omega) + N_u(\tau,\omega)}{S_k(\tau,\omega) + N_o(\tau,\omega)} = \frac{u_k + \varepsilon_{us}}{1 + \varepsilon_{os}} \quad (14)$$

where $\varepsilon_{us} = N_u(\tau,\omega) / S_k(\tau,\omega), \varepsilon_{os} = N_o(\tau,\omega) / S_k(\tau,\omega)$. The TF index (τ, ω) has been dropped for simplifying the notation. If signal-to-noise ratio is high (additive noise is much smaller than the sources), from (14) we have

$$I_{uo}(\tau, \omega) \approx u_k + \varepsilon_{uo}(\tau, \omega) \quad (15)$$

Similar derivations can be accordingly drawn to other ISDRs as follows

$$I_{vo}(\tau, \omega) = v_k + \varepsilon_{vo}(\tau, \omega) \quad (16)$$

$$I_{wo}(\tau, \omega) = w_k + \varepsilon_{wo}(\tau, \omega) \quad (17)$$

where $\varepsilon_{uo}(\tau, \omega), \varepsilon_{vo}(\tau, \omega)$ and $\varepsilon_{wo}(\tau, \omega)$ can be viewed as the ISDR modeling error with zero mean introduced by the additive noise. From (15) to (17), it's clear to see that the ISDRs $I_{uo}(\tau, \omega)$, $I_{vo}(\tau, \omega)$ and $I_{wo}(\tau, \omega)$ can be viewed as random variables in TF domain with the mean of u_k, v_k and w_k, respectively. Specifically, DOA estimation task can be achieved by clustering the ISDRs corresponding to all TF points and estimate the *K* clusters centered at (u_k, v_k, w_k), $k = 1, ..., K$. Then the estimated $(\hat{u}_k, \hat{v}_k, \hat{w}_k)$ can be further used to estimate the DOA of the *k*-th source by

$$\hat{\theta}_k = \cos^{-1} \hat{w}_k, \quad \hat{\phi}_k = \tan^{-1} \hat{v}_k / \hat{u}_k \quad (18)$$

Research shows that our proposed multisource DOA estimation method using a single AVS performs perfectly under noiseless condition and performs well when the SNR is high and the TF sparsity is satisfied. When the SNR is low, it is easy to understand that the ISDR error in (15) to (17) will corrupt the ISDRs. Moreover, the sparsity may not be present at some TF points. Under such conditions, the relation between the ISDRs and the direction-cosines shown in (15) to (17) may not hold and the proposed DOA estimation approach may fail. Hence it is a straightforward strategy that a robust DOA estimation algorithm should be developed by only using the ISDRs of TF points with high local SNR and good TF sparsity.

2.3. Proposed Robust ISDR-DOA Estimation Algorithm

In this section, the method to extract the TF points with high local SNR (HLSNR-points) will be addressed first. Then the clustering method will be introduced to estimate the cluster centers of ISDRs. Finally, the robust ISDR-DOA estimation algorithm will be developed and summarized.

2.3.1. High Local SNR ISDRs Extraction in TF Domain

According to the analysis above, the selection of the HLSNR-points is important for achieving robustness and good performance of the proposed DOA estimation method. Sinusoidal tracks extraction (SinTrE) methods [3][8] are effective approaches to extract HLSNR-points for speech signals where the harmonic structures of the speech and high energy property have been fully explored. In our development, we adopt the method proposed in [8] to select the trustable ISDRs associated with HLSNR-points. By carefully examining the SinTrE method, we noted that if there are multiple speech sources with similar power level present at a TF point or nearby, the performance of SinTrE method will degrade since the interaction between the sources causes the fluctuation of the mixed spectrum, which frequently results in no sinusoidal tracks in this TF domain. However, the SinTrE method works well when the STFT of signals has good sparsity in the TF domain.

2.3.2. Clustering

When the HLSNR-points have been extracted, the effective and robust clustering algorithm should be applied to estimate the cluster centers. The histogram method [1][5] is the simplest solution. However, the selection of the low pass filter bandwidth is quite tricky. Improper bandwidth will cause large DOA estimation error. Evaluation of the sophisticated clustering methods such as K-means or EM shows that the DOA estimation would be biased. This is because the clustered centers estimated by these methods will be affected greatly by the possible unreliable HLSNR-points. To achieve a better clustering result, kernel density estimation (KDE) method [15] is adopted in our study. The KDE can be viewed as a generalization of the histogram with improved statistical properties. For single AVS DOA estimation method, from (2), it is noted that there is the joint relationship $u_k^2 + v_k^2 + w_k^2 = 1$. Therefore, the clustering problem can be further formulated as a bivariate KDE problem, where the KDE gives the joint probability density function (PDF) of ISDRs associated with the extracted HLSNR-points. As a result, the locations of first K peaks of the estimated joint PDF correspond to the K pairs of direction-cosines. Since the PDF estimated by the KDE has reflected the statistical property of a local TF area, the K peaks of the estimated joint PDF will be less affected by the unreliable HLSNR-points.

2.3.3 AVS-based ISDR-DOA Estimation Algorithm

Following the derivation and discussion above, the proposed robust DOA estimation algorithm by using a single AVS unit, named here as the AVS-based ISDR DOA estimation algorithm (ISDR-DOA in short) is summarized as follows.

1) Segment the AVS output data by N-length Hamming window;
2) Calculate the DFT of the segments;
3) Extract the sinusoidal tracks for all four sensors to form a joint sinusoidal track set (HLSNR-points) ;
4) Calculate the ISDRs between the *u*-, *v*-, *w*- and *o*-sensor by (11)-(13) and select the ISDRs according to the HLSNR-points computed in step 3);
5) Calculate the joint PDF over either two of the three selected ISDRs using the bivariate KDE method [15];
6) Locate first K peaks in the joint PDF and calculate DOAs using the estimated direction-cosines via (18).

3. EXPERIMENTAL RESULTS

In this section, three simulations and one experiment have been conducted to evaluate the performance of our proposed ISDR-DOA algorithm. For simulation studies, the simulation parameters are set as follows: 1) 8kHz sampling rate and 3 seconds speech are used; 2) The 512-point DFT and 256-point Hamming window with 60% overlapping are adopted. The statistical independent additive Gaussian white noise is considered for each AVS sensor.

Experiment 1*: Multisource DOA Estimation Capability*

This simulation has been carried out to evaluate the multisource DOA estimation capability of our proposed ISDR-DOA algorithm under noise condition, where the SNR is fixed to be 20dB. In Table I, the "True DOA" indicates the ideal DOA's of the seven speech sources; "Estimated" means the estimated DOA by our proposed ISDR-DOA algorithm; "DOA error" represents the absolute estimation errors between "True DOA" and "Estimated" DOA. For the purpose of the visualization, the estimated joint PDF of the direction-cosines has been transformed into the angular space and is shown in Fig. 1. It is clear and very encouraging to see that the ISDR-DOA algorithm with single AVS unit has the ability to accurately estimate all seven sources simultaneously, which is one of the attractive properties of the proposed method.

Experiment 2*: DOA estimation Accuracy*

This simulation has been performed to test the DOA estimation accuracy of our proposed ISDR-DOA algorithm for different DOA's. The GMDA-Laplace method [3] has been taken for performance comparison since this method was developed using two microphones based on the TF sparsity as well. The setups are as follows: 1) two microphones are placed along *z*-axis with 8cm spacing; 2) Two speech sources, denoted as *S1* and *S2*, are considered. *S1* is located in a fixed position at (45°, 45°), and the location of *S2* changes from (50°, 45°) to (80°, 45°) ; 3) The signal to noise ratio (SNR) is set to be 20dB. The root mean squared error (RMSE) is used as the performance measure, and the results are obtained by 100 independent trials. The simulation results are plotted in Fig. 2. It is clear to see that the DOA estimation accuracy of our proposed ISDR-DOA

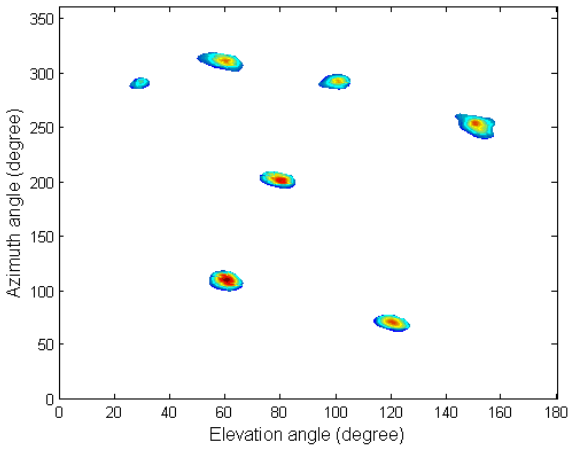

Fig. 1. Illustration of seven source DOA estimation

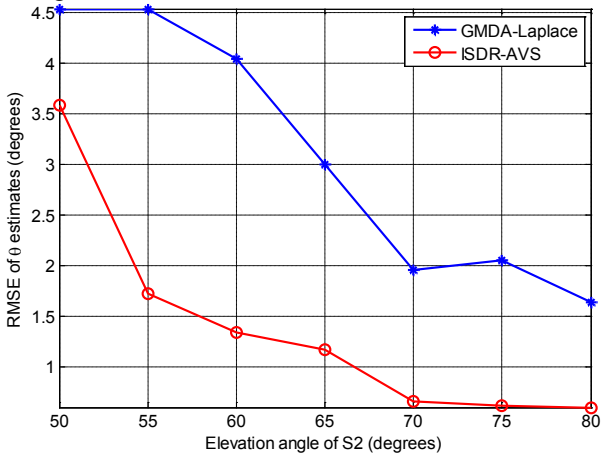

Fig. 2 RMSE *versus* different DOA's

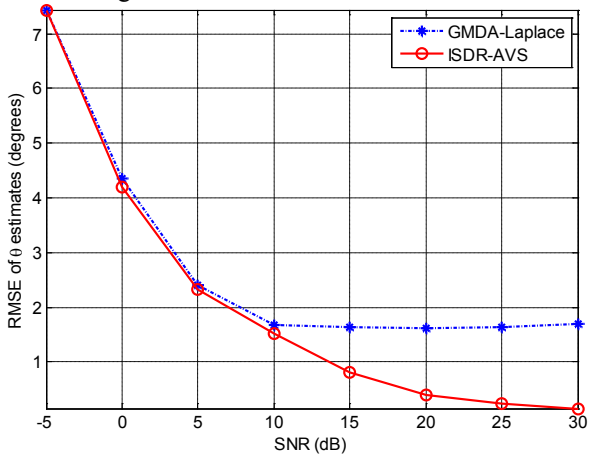

Fig. 3 RMSE *versus* SNR

is superior to that of the GMDA-Laplace under all source separations. As expected, the estimation accuracy of both algorithms increases as the source separation becomes larger.

Experiment 3*: Robustness of the DOA estimation.*

This simulation aimed at evaluating the influence of the SNR on the performance of DOA estimation algorithms. The simulation conditions are the same as those in Experiment 2 except the *S1* and *S2* is located at (60°, 45°) and (135°, 45°) respectively and the *SNR* varies from -5dB to 30dB. Experiment results are shown in Fig. 3. It is noted that our proposed ISDR-DOA algorithm performs comparable to that of the GMDA-Laplace when the SNR is smaller than 10dB and outperforms the latter when the SNR is larger than 10dB. It can be seen that the performance of the ISDR-DOA algorithm is of high accuracy when SNR is higher than 15dB (RMSE less than 1degree).

Experiment 4: *DOA Estimation Based on Recorded Data*

In this experiment, we conducted the DOA estimation by using the recorded data from Ritz's lab [14]. The AVS has been built by two Knowles NR-3158 pressure gradient sensors (*u*- and *v*-sensors) and one Knowles EK-3132 sensor (*o*-sensor). Recordings were made of 10 different speech sentences from the IEEE speech corpus [15] in a reverberant room with of RT_{60}=30ms and background noise includes the noise from computer servers and air conditioning. The sampling rate was 48kHz and then downsampled to 16kHz for DOA estimation. The speakers were placed in the front of the AVS at a distance of 1m. *S1* was located at a fixed position (0°, 0°), while *S2* moved from (0°, 15°) to (0°, 75°). The average DOA estimations over all examples of *S2* are listed in the second row of Table II, while the true DOA's listed in the first row. From Table II, we can see that our proposed ISDR-DOA algorithm provides high DOA estimation accuracy with real recorded data. These preliminary experimental results further validate the assumptions and derivation of ISDR-DOA method.

4. CONCLUSION

In this paper, based on the assumption of the speech spectrum sparsity, we developed a relation between the inter-sensor data ratio (ISDR's) of single AVS and the DOA's of the spatial sources. With the high local SNR TF point extraction and KDE clustering techniques, a robust ISDR-DOA estimation algorithm for multiple speech sources using a single AVS has been developed and evaluated. Simulation results show that the proposed ISDR-DOA algorithm illustrates a desirable performance, such as capability to estimate up to seven DOAs simultaneously, robustness to additive noise, computational efficiency, high DOA estimation accuracy when SNR larger than 15dB. It is expected that the proposed ISDR-DOA method may find wide applications in portable devices.

Table I. Estimation results of ISDR-DOA Algorithm for Seven sources (°)

True DOA	(30, 290)	(60, 110)	(100, 290)	(120, 70)
Estimated	(30.59,289.4)	(59.48, 111.4)	(99.84, 291.2)	(120.3, 70.21)
DOA Error	(0.59,0.6)	(0.52,1.4)	(0.16,1.2)	(0.3,0.21)
True DOA	(150, 250)	(60, 310)	(80,200)	
Estimated	(150.7, 251.9)	(60.48, 310)	(78.91,201.3)	
DOA Error	(0.7,1.9)	(0.48,0)	(1.09,1.3)	

Table II. DOA Estimation Results of AVS

True DOA (°)	15	30	45	60	75
Measured (°)	14.96	31.46	47.85	56.41	66.44

5. REFERENCES

[1] C. Liu, B. C. Wheeler, etc., "Localization of multiple sound sources with two microphones," J*ournal of the Acoustical Society of America*, vol. 108, pp. 1888-1905, 2000.

[2] M. I. Mandel and D. P. W. Ellis, "EM localization and separation using interaural level and phase cues," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 275-278, 2007.

[3] W. Zhang and B. D. Rao, "A two microphone-based approach for source localization of multiple speech sources," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 1913-1928, 2010.

[4] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-Frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, pp. 1830-1847, 2004.

[5] M. Matsuo, Y. Hioka, and N. Hamada, "Estimating DOA of multiple speech signals by improved histogram mapping method," *International Workshop on Acoustic Echo and Noise Control*, pp. 129-132, 2005.

[6] S. Araki, H. Sawada, R. Mukai, S. Makino, "DOA estimation for multiple sparse sources with normalized observation vector clustering," *IEEE International Conference on Acoustics, Speech and Signal Processing*, V33-V36, 2006.

[7] M. Cobos, J. J. Lopez, and D. Martinez, "A sparsity-based approach to 3D binaural sound synthesis using time-frequency array processing," *EURASIP J. Adv. Signal Process*, vol. 2010, pp. 1-13, 2010.

[8] M. Ren and Y. X. Zou, "A Novel Multiple Sparse Source Localization Using Triangular Pyramid Microphone Array," *IEEE Signal Processing Letters*, Vol. 19, No. 2, p83-86, February, 2012.

[9] M. Hawkes, A. Nehorai., "Acoustic vector sensor beamforming and Capon direction estimation," *IEEE Trans. Signal Process.*, vol.46, no.9, pp.2291-2304, Sep. 1998.

[10] P. Tichavsky, K. T. Wong and M. D. Zoltowski, "Near-Field/Far-Field Azimuth & Elevation Angle Estimation *Using* a Single Vector-Hydrophone," *IEEE Trans. Signal Process.*, vol. 49, no. 11, pp. 2498-2510, Nov. 2001.

[11] M. Shujau, C. H. Ritz, and I. S. Burnett, "Using in-air Acoustic Vector Sensors for tracking moving speakers". *the 4th International Conference on Signal Processing and Communication Systems (ICSPCS)*, pp. 1-5, 2010.

[12] B. LI and Y. X. ZOU, "Improved DOA estimation with acoustic vector sensor arrays using spatial sparsity and subarray manifold," *IEEE International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, March 25-30, 2012.

[13] P. K. T. Wu, C. Jin, and A. Kan, "A Multi-Microphone Speech Enhancement Algorithm Tested Using Acoustic Vector Sensors," *International Workshop on Acoustic Echo and Noise Control*, 2010.

[14] C. H. Ritz, and I. S. Burnett, "Separation of speech sources using an acoustic vector sensor," *2011 IEEE internatinal Workshop on Multimedia Signal Processing (MMSP)*, Hanzhou, China, 2011.

[15] Z. I. Botev, J. F. Grotowski and D.P. Kroese, "Kernel density estimation via diffusion". *Annals of Statistics*, vol. 38, no. 5, pp. 2916–2957, 2010.

[16] IEEE Subcommittee (1969). "IEEE recommended practice for speech quality measurements". *IEEE Trans. Audio and Electro-acoustics*, AU-17(3), 225-246.