Learning a Robust DOA Estimation Model with Acoustic Vector Sensor Cues

Yuexian Zou[1*], Rongzhi Gu[1,2], Disong Wang[1], Aimin Jiang[2] and Christian H. Ritz[3]
[1] ADSPLAB/Intelligent Lab, School of ECE, Peking University, Shenzhen,518055, China
[2] College of Internet of Things Engineering, HoHai University, Changzhou, China
[3] School of Electrical, Computer, and Telecommunications Engineering, University of Wollongong, Australia
*E-mail: zouyx@pkusz.edu.cn

***Abstract*— Accurate and robust Direction of Arrival (DOA) estimation with small microphone arrays is gaining an increasing demand in service robotics and smart home applications. Classic non-learning DOA estimation methods show unsatisfactory performance under low SNR or high reverberation conditions. Meanwhile, some research outcomes illustrate that learning methods with Neural Networks (NN) ask for careful array element quantity or layout regulation which is impractical for many applications. In order to obtain robust DOA estimation with small arrays, taking the learning ability of Deep Neural Networks (DNN), we propose to form the training pairs by using Acoustic Vector Sensor – Direction of Arrival (AVS-DOA) cues and its counterpart DOA which can be simulated under different SNR and reverberation conditions. Then DNN-based DOA model is trained accordingly and the performance of the model has been fully investigated with different activation functions, network structures and dropout rates. With the cross-validation process, the model performing best experimentally is selected as the optimal DOA model. Experimental results validate the effectiveness of our DNN based DOA model which outperforms the non-learning method, especially under poor acoustic conditions.**

I. INTRODUCTION

Since the burgeoning of artificial intelligence, direction of arrival (DOA) estimation of a single spatial speech source using small microphone arrays has attracted a considerable amount of attention in many applications, such as service robotics, smart home applications, video conference, etc. However, conventional non-learning DOA estimation methods [1] suffer from various kinds of noise and high reverberation, showing a reliance on the cleanness of the acoustic environment. Meanwhile, learning methods with NN try to learn the mapping between the inputs and the DOAs in challenging conditions for robust DOA estimation [2] - [3]. Nevertheless, they tend to claim high demand for array geometry and array element quantity.

Currently, the acoustic vector sensor (AVS) has been exploited as an ideal candidate for DOA estimation with small physical sensor size constraint. Obviously, AVS has appealing characteristics of strong direction sensitivity and large acquisition quantity. Research shows that the AVS has a superior ability to yield higher resolution and accuracy of DOA estimation with certain proposed research scenarios [4].

Since Nehorai found that the AVS outperforms the pressure-sensor in DOA estimation and proposed DOA measurement method based on AVS in 1994 [5], many conventional DOA estimation methods have been investigated for the AVS. Moreover, high-order statistics methods proposed by Y. H. Jin [6] and A. Agarwal [7], both give satisfactory experimental results for DOA estimation using a single AVS. However, these existing methods usually call for complicated data processing and adjustment of hyper-parameters for different application conditions, which lead to high computational cost as well as inconvenience in real applications. Exploring the time-frequency (TF) sparsity and fully investigating the properties of signals captured by an AVS, our previous work derived DOA estimation using inter-sensor data ratios (ISDRs) [8], where the high local SNR TF points have been determined and the ISDRs are clustered with kernel density estimation (KDE) techniques to give the DOA estimation. Simulation results validate its merits in terms of less computational cost and relatively high accuracy under noiseless or moderate noise conditions. However, experimental results also indicate that this method generates biased DOA estimation results under strong noise and reverberation conditions, which hinders its applications in real environments.

Following the discussions above, we may conclude that so far accurate and robust DOA estimation with a single AVS in adverse environments is still a challenging task. In this paper, from data-driven and machine learning perspectives, we present a novel DNN-based single source DOA estimation approach in noisy and reverberant acoustic environment via using AVS-DOA cues. With the relation between the DOAs and the AVS cues, a large scale training data is simulated using different signal settings and various acoustic conditions. To evaluate our method, extensive experiments have been carried out which give encouraging results, especially under adverse acoustic conditions.

II. AVS DATA MODEL

In this paper, we consider utilizing an AVS with 3 channels incorporating one omnidirectional sensor and two orthogonally oriented directional sensors [4], which are termed as *o*-sensor, *u*-sensor and *v*-sensor respectively. The DOA of the speech source is denoted as (θ, ϕ) where the elevation angle $\theta \in (0°, 180°)$ and the azimuth angle $\phi \in [0°, 360°)$. In this paper θ is fixed at 90° for simplicity's sake. Then the manifold vector for the AVS is given by [8]

$$\mathbf{a} = [u \; v \; o]^T = [\cos\phi \; \sin\phi \; 1]^T \tag{1}$$

where the elements *u, v* correspond to the *x-, y-* axis, respectively, and $\cos\phi, \sin\phi$ are derived according to coordinates geometry.

Then, in practical acoustic environments with noise and reverberation, the output of the calibrated AVS at discrete time *t* can be expressed as follows [9]:

$$\mathbf{y}(t)=\mathbf{h}(t)*s(t)+\mathbf{n}(t) \quad (2)$$

where $\mathbf{y}(t)=[y_u(t), y_v(t), y_o(t)]^T$ denotes the output signal of *u*-, *v*-, *o*- sensor, respectively, $\mathbf{n}(t)=[n_u(t), n_v(t), n_o(t)]^T$ is additive noises at each sensor, $\mathbf{h}(t)=[h_u(t), h_v(t), h_o(t)]^T$ is the impulse response vector from speech source $s(t)$ to the corresponding sensor. Taking the STFT of (2) gives [9]

$$\mathbf{Y}(k,m)=\mathbf{H}(k)S(k,m)+\mathbf{N}(k,m) \quad (3)$$

where $S(k, m)$ is STFT of $s(t)$, $\mathbf{Y}(k, m)=[Y_u(k, m), Y_v(k, m), Y_o(k, m)]^T$, $\mathbf{N}(k, m)=[N_u(k, m), N_v(k, m), N_o(k, m)]^T$ are the STFT of $y(t)$ and $n(t)$ respectively. In addition, $\mathbf{H}(k)=[H_u(k), H_v(k), H_o(k)]^T$ can be represented as

$$\mathbf{H}(k)=\mathbf{H}^d(k)+\mathbf{H}^r(k) \quad (4)$$

where $\mathbf{H}^d(k)$ and $\mathbf{H}^r(k)$ are the direct and reflection components of the room impulse response $\mathbf{H}(k)$, and can be expressed as

$$\mathbf{H}^d(k)=e^{-j\omega_k\tau}\mathbf{a},\ \mathbf{H}^r(k)=\sum\nolimits_j \lambda^j e^{-j\omega_k\tau^j}\mathbf{a}^j \quad (5)$$

where τ denotes the direct-path time delay, ω_k is the kth discrete angular frequency, τ^j and λ^j are the jth path time delay of the reflection and attenuation due to absorption at surfaces of the room. Note that the attenuation of $\mathbf{H}^d(k)$ is normalized to 1 for simplicity. Substituting (5) into (3), we have

$$Y_u(k,m)=S(k,m)[e^{-j\omega_k\tau}u+\sum\nolimits_j \lambda^j e^{-j\omega_k\tau^j}u^j]+N_u(k,m) \quad (6)$$

$$Y_v(k,m)=S(k,m)[e^{-j\omega_k\tau}v+\sum\nolimits_j \lambda^j e^{-j\omega_k\tau^j}v^j]+N_v(k,m) \quad (7)$$

$$Y_o(k,m)=S(k,m)[e^{-j\omega_k\tau}+\sum\nolimits_j \lambda^j e^{-j\omega_k\tau^j}]+N_o(k,m) \quad (8)$$

From (6) to (8), we can clearly see that $Y_u(k, m)$ and $Y_v(k, m)$ comprise the DOA cues (u, v) and other signal components due to additive noise as well as reverberation components.

Through the careful analysis in [8], the inter-sensor data ratios (ISDRs) at speech dominated TF points (TFPs) are derived and proved to be effective cues for DOA estimation. Specifically, the ISDR between *u*- and *o*-sensor is defined as follows

$$\mathrm{ISDR}_{uo}(k,m)=Y_u(k,m)/Y_o(k,m) \quad (9)$$

Substituting (6) and (8) into (9) gives

$$\begin{aligned}\mathrm{ISDR}_{uo}(k,m)&=\frac{S(k,m)[e^{-j\omega_k\tau}u+\sum_j \lambda^j e^{-j\omega_k\tau^j}u^j]+N_u(k,m)}{S(k,m)[e^{-j\omega_k\tau}+\sum_j \lambda^j e^{-j\omega_k\tau^j}]+N_o(k,m)}\\&=(u+\varepsilon_u)/(1+\varepsilon_o)\end{aligned} \quad (10)$$

where ε_u and ε_o are given by

$$\varepsilon_u=\sum\nolimits_j \lambda^j e^{-j\omega_k\tau^{j-1}}u^j+\left(N_u(k,m)\right)/\left(S(k,m)e^{-j\omega_k\tau}\right) \quad (11)$$

$$\varepsilon_o=\sum\nolimits_j \lambda^j e^{-j\omega_k\tau^{j-1}}+\left(N_o(k,m)/\left(S(k,m)e^{-j\omega_k\tau}\right)\right) \quad (12)$$

To simplify the expression, the TF index (k, m) is omitted. According to Taylor's formula, when the convergence condition $|\varepsilon_o| < 1$ is satisfied (when the SNR is high), (9) can be rewritten as

$$\mathrm{ISDR}_{uo}(k,m)=(u+\varepsilon_u)(1-\varepsilon_o+\varepsilon_o^2-\varepsilon_o^3+\cdots) \quad (13)$$

Neglecting high order terms, we have

$$\mathrm{ISDR}_{uo}(k,m)\approx u+\varepsilon_{uo} \quad (14)$$

where $\varepsilon_{uo}=\varepsilon_u-u\varepsilon_o-\varepsilon_u\varepsilon_o$ can be regarded as ISDR error bought by noise and reverberation.

Symmetric derivation of ISDR between the *v*-sensor and *o*-sensor can be obtained as follows

$$\mathrm{ISDR}_{vo}(k,m)\approx v+\varepsilon_{vo} \quad (15)$$

Obviously, the errors ε_{uo} and ε_{vo} go to zero under noiseless and no reverberation condition. Under noisy and reverberant conditions, the ISDRs at speech dominated TFPs (ε_{uo} and ε_{vo} are of small values) contain more direction information (u, v), which enables the ISDRs to be efficient AVS-DOA cues for accurate DOA estimation.

III Proposed DNN-Based Method

In this section, we firstly describe the details of the extraction of AVS-DOA cues. Then, our proposed method, termed as ISDR-DNN in short, is presented, where the DOA estimation task is formulated as a classification problem. A robust DOA estimation model is learned by DNN which actually maps the AVS-DOA cues to the corresponding DOA in a supervision manner.

A. AVS-DOA cues extraction

Through the analysis presented in Section II, it can be deduced that ISDRs at speech dominated TFPs can be considered as DOA cues, which provide DOA information. Therefore, a robust and accurate method to extract speech dominated TFPs becomes the key. In this study, we employ the sinusoidal tracks extraction (SinTrE) method [10] since it exploits the harmonic structure of speech to find high local SNR regions. SinTrE has been proved to be a reasonable good and reliable method to extract speech dominated TFPs. However, experimental results show that the amount of speech dominated TFPs extracted by SinTrE varies due to several factors, such as SNR levels, reverberation time and DOA of the speaker. Hence, a further process is required to extract reliable and the same amount of the speech dominated TFPs for different signal conditions. Based on experimental observations, we use 200 ISDRs at extracted speech dominated TFPs for DOA estimation. Specifically, assuming that W speech dominated TFPs have been extracted, then W ISDRs can be sorted in a descending order as follow

$$\{r_1,r_2,...,r_W\},(r_1\ge r_2\ge ...\ge r_W) \quad (16)$$

where r_w is the wth ISDR. If W is no less than 200, we keep the first 200 ISDRs. If W is less than 200, the number of ISDRs is extended to 200. In detail, we denote a new sequence $\{R_{i_1}, R_{i_2}, \ldots, R_{i_W}\}$ where

$$R_{i_w}=r_w \quad \text{and} \quad i_w=<200W/w>,\ \ 1\le w\le W \quad (17)$$

where $<\cdot>$ represents the round operator. Then we perform the quadratic fit on the $\{(i_w, R_{i_w}), 1\le w\le W\}$ to obtain the quadratic function $f(\cdot)$. Supposing that the $\{j_1, j_2, \ldots, j_{200-W}\}$ is the difference set between $\{1, 2, \ldots, 200\}$ and $\{i_1, i_2, \ldots, i_W\}$. Then, the another 200-$W$ ISDRs are calculated as $\{f(j_p), 1\le p\le 200-W\}$, which are added to the original W ISDRs to form the 200-dimensional ISDRs.

To demonstrate the efficiency of ISDRs extracted as the input feature, a toy example is shown in Fig. 1, where we present the ISDRs patterns versus speech dominated TFPs at different DOAs (0°, 30°, 60°, 90°) and with different SNRs (from -5dB to 25dB with 5dB interval). The reverberation time is fixed at 0.15s. Similarly, ISDR patterns versus the speech dominated TFPs at different DOAs with the different

reverberation time (0.15s, 0.30s, 0.45s, 0.60s and 0.75s) are illustrated in Fig. 2 and SNR is fixed at 20dB. From Fig. 1 and Fig. 2, intuitively, for a given DOA, we observe similar patterns in most cases although they are associated with different SNR levels and reverberation conditions. These results indicate that the ISDRs at the speech dominated TFPs can be taken as efficient AVS-DOA cues for DOA estimation. Besides, to fully exploit the ISDRs, we construct 400-dimensional AVS-DOA cues by cascading the 200-dimensional $ISDR_{uo}$ and 200-dimensional $ISDR_{vo}$. For presentation consistency, in the following context, these 400-dimensional vectors are termed as AVS-DOA cues and will be used in the following context.

B. Learning a DOA estimation model

In this study, we try to solve the DOA estimation from a learning perspective, where the idea is to learn a mapping between the AVS-DOA cues and the DOA learnt using the DNN. Therefore, we design the DOA estimation model by using AVS-DOA cues as the input of the DNN and the ground truth DOAs as the training data, as illustrated in Fig. 3.

In the training stage, to improve the generalization ability of the DNN for tackling a variety of adverse environments, we create a training dataset of AVS-DOA cues extracted under different noisy and reverberant conditions (details are given in Sect. IV), and the corresponding DOAs are used as the ground truth. With the training dataset, the DNN is trained and the DOA estimation model is obtained.

In the testing stage, following the same procedure, the AVS-DOA cues are firstly extracted, then decoded by the

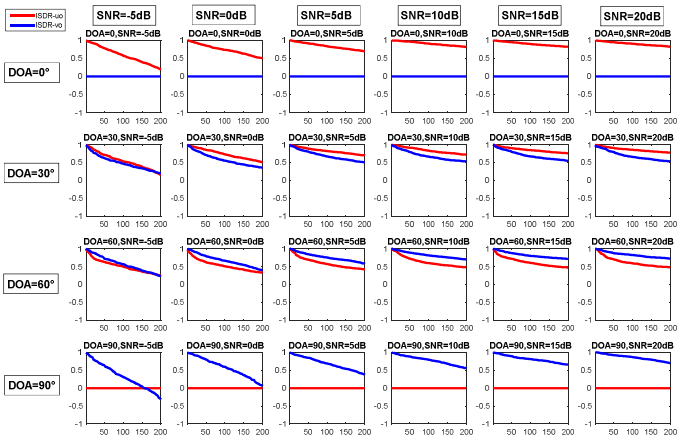

Fig. 1 ISDRs patterns of different DOA with different SNR levels

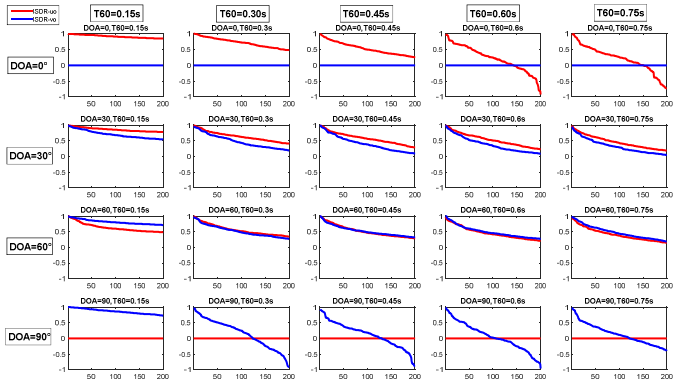

Fig. 2 ISDRs patterns for different DOA with different reverberation time

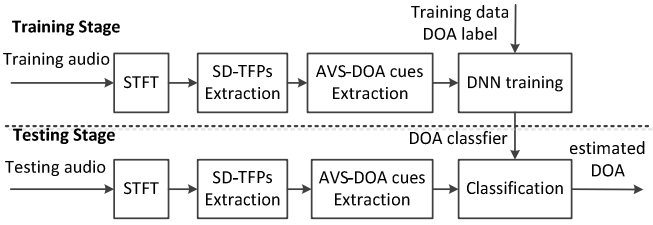

Fig. 3 Diagram of ISDR-DNN

trained DNN model to predict the DOA.

III. Simulations and Analysis

To create the dataset for training the DNN, the received signal $\mathbf{y}(t)$ of the AVS is generated according to (2) where $\mathbf{h}(t)$ is simulated following the method proposed in [11], and $\mathbf{n}(t)$ is of Gaussian distribution. The signals are sampled at 8kHz. 256-point Hamming window with 50% overlapping is used for the STFT calculation. To obtain ISDRs in a variety of conditions, we simulate $\mathbf{y}(t)$ with different noise and reverberation levels. The detailed parameters used are shown in Table I. 50 sentences randomly selected from TIMIT [12] are used as the speech sources, and each sentence is repeatedly used for different experimental settings. We divide the dataset into training, verification and test dataset with the percentage of 70%, 20% and 10%, respectively. To verify the effectiveness of our proposed method, the non-learning method ISDRWO [8] is used as the baseline, since this method also uses ISDRs as AVS-DOA cues for DOA estimation. Besides, the absolute average error (AAE) and root mean square error (RMSE) are employed as the performance evaluation measures of DOA estimation.

A. DNN Model comparison and selection

Note that the performance of DNN is substantially influenced by the hyper-parameters. In order to obtain a proper model, we tried various hyper-parameters, including the activation functions, dropout rates, and the dimension of hidden layers with the number of hidden layers fixed at 2. The models are trained with the training dataset and evaluated by the verification dataset [13], and the results are shown in Table □. We can see that the best performance is given by the model with a 400-256-256-360 network structure, Leaky ReLU activation function and dropout rate of 0.2, which is selected as the optimal model, as shown in Fig. 4.

B. DOA estimation accuracy versus different SNRs

This simulation is carried out on MATLAB to evaluate robustness of our DNN-based DOA estimation model under different noise levels. The room is set to be 7m×5m×3m with an AVS located in the center and the distance between the source and AVS fixed as 2m. Room reverberation time is 0.15s.

We examine the performances of ISDR-DNN and

TABLE I
CONFIGURATIONS USED FOR GENERATING DATA

Speech	Sentences from TIMIT Dataset
Room size(m)	small (7*5*3), medium (12*10*3), large (17*15*3)
Distance(m)	near (1), far (2 for small, 4 for medium, 6.5 for large)
SNR(dB)	-5 to 25 with 5 step
T60(s)	0.15 to 0.75 with 0.15 step

ISDRWO with SNRs ranging from -5dB to 25dB with 5dB increments. The results are illustrated in Fig. 5. It is observed that as the noise level increases, the performance of ISDRWO degrades dramatically, especially when the SNR is under 5dB. By contrast, the performance of ISDR-DNN comparatively remains steady. Moreover, the ISDR-DNN exhibits a better accuracy than ISDRWO at the same SNR level, which demonstrates the superiority of our proposed method over ISDRWO under different noisy conditions.

C. *DOA estimation accuracy versus different reverberation*

This simulation aims at evaluating the influence of reverberation on the performance of DOA estimation. We used the previous settings except that the SNR is fixed at 20dB and the reverberation time ranges from 0.15s to 0.75s with 0.15s increments. Simulation results are shown in Fig. 6. As expected, the performance of two algorithms deteriorates with the increasing reverberation level. However, the performance of ISDR-DNN is still of higher accuracy than that of ISDRWO. Our purposed model can estimate source direction with less than 2° error for reverberation time of 0.15s. Overall, ISDR-DNN has smaller estimation errors and is more stable under different reverberation conditions, showing the efficiency and robustness of our proposed method.

IV. CONCLUSIONS

In this paper, by analyzing the effectiveness of ISDRs at speech dominated TFPs as AVS-DOA cues, we present a DOA estimation from a learning perspective with a single AVS, where the DNN is employed for mapping the AVS-DOA cues to the corresponding DOA. By adjusting the hyper-parameters of the DNN with the training dataset generated in various noisy and reverberant conditions, we select the optimal DNN model for DOA estimation, which has the lower absolute average error and RMSE as compared to ISDRWO, a non-learning method. For the future work, we expect improved results with larger training datasets or better DNN structure and settings.

TABLE □
ABSOLUTE AVERAGE ERROR OF TRAINED MODELS WITH DIFFERENT TRAINING PARAMETERS ON VALIDATION DATASET

Activation Function	Dropout Rate	Number of units per hidden layer			
		256	400	600	800
ReLU	0.10	3.77	3.84	3.78	3.81
	0.15	3.80	3.87	3.81	3.79
	0.20	3.70	3.87	3.80	3.88
	0.25	3.79	3.93	3.89	3.79
Leaky ReLU	0.10	3.71	3.73	3.71	3.72
	0.15	3.72	3.73	3.73	3.77
	0.20	**3.60**	3.73	3.87	3.77
	0.25	3.81	3.85	3.83	3.85
PReLU	0.10	3.79	3.81	3.69	3.77
	0.15	3.78	3.77	3.76	3.72
	0.20	3.69	3.82	3.84	3.90
	0.25	3.69	3.84	3.90	3.94

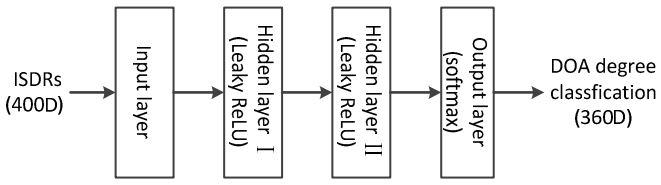

Fig. 4 The structure of DNN-based DOA estimation model

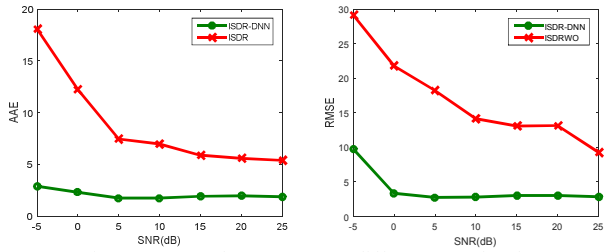

Fig. 5 AAE and RMSE versus different SNR levels

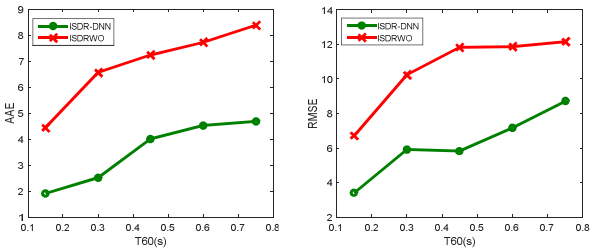

Fig. 6 AAE and RMSE versus different reverberation time

ACKNOWLEDGEMENT

This project was partially supported by Shenzhen Science & Technology Fundamental Research Programs (No: JCYJ 20170306165153653) & Shenzhen Key Laboratory for Intelligent Multimedia and Virtual Reality (No: ZDSYS 201703031405467).

REFERENCES

[1] C. Knapp and G. Carter, "The generalizaed correlation method for estimation of time delay", *IEEE ICASSP*, pp. 320-327, 2003.
[2] M. Aktas, T. Akgun, D. Buyukaydin and H. Ozkan, "Acoustic direction finding under high reverberation," *IEEE Signal Processing and Communications Applications Conference*, pp. 1533-1536, 2015.
[3] X. Xiao, S. Zhao, X. Zhong, et al., "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," *IEEE ICASSP*, pp. 2814-2818, 2015.
[4] J. W. Cao, J. Liu, J. Z. Wang and X. P. Lai. "Acoustic vector sensor: reviews and future perspectives," *Iet Signal Processing*, pp. 1-9, 2017.
[5] A. Nehorai and E. Paldi, "Acoustic vector-sensor array processing", *IEEE Transactions on Signal Processing*, vol. 9, pp. 2481-2491, 1994.
[6] Y. H. Jin and Y. X. Zou, "Robust speaker DOA estimation with single AVS in bispectrum domain," *IEEE ICASSP*, pp. 2814-2818, 2015.
[7] A. Agarwal, A. Kumar and M. Agrawal, "Higher order statistics based Direction of Arrival estimation with single Acoustic Vector Sensor in under-determined case," *IEEE Oceans*, pp. 1-10, 2015.
[8] Y. X. Zou, W. Shi, B. Li, et al. "Multisource DOA estimation based on time-frequency sparsity and joint inter-sensor data ratio with single acoustic vector sensor", *IEEE ICASSP*, pp. 4011-4015, 2013.
[9] K. W, V. G. Reju and A. W. H. Khong, "Multi-source direction-of-arrival estimation in a reverberant environment using single acoustic vector sensor" IEEE ICASSP, pp. 444-448, 2015.
[10] Mcaulay, J. Robert and T. F. Qualtieri, "Speech analysis/Synthesis based on a sinusoidal representation," *IEEE ICASSP*, pp. 744-754, 1986.
[11] E. Habets, "Room Impulse Response Generator," *Technische Universiteit Eindhoven. Tech. Rep*, pp. 1-21, 2010.
[12] J. S. Garofolo, "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," *National Institute of Standards and Technology (NIST). Gaithersburgh. MD*, vol. 107, 1988.
[13] F. Chollet, "Keras: Theano-based deep learning library," *https://github.com/fchollet/keras*, 2015.