

# A LEARNING-BASED APPROACH TO DIRECTION OF ARRIVAL ESTIMATION IN NOISY AND REVERBERANT ENVIRONMENTS

Xiong Xiao<sup>1</sup>, Shengkui Zhao<sup>2</sup>, Xionghu Zhong<sup>3</sup>, Douglas L. Jones<sup>2</sup>, Eng Siong Chng<sup>1,3</sup>, Haizhou Li<sup>1,3,4</sup>

<sup>1</sup>Temasek Lab@NTU, Nanyang Technological University, Singapore

<sup>2</sup>Advanced Digital Sciences Center, Singapore

<sup>3</sup>School of Computer Engineering, Nanyang Technological University, Singapore

<sup>4</sup>Department of Human Language Technology, Institute for Infocomm Research, Singapore

xiaoxiong@ntu.edu.sg, shengkui.zhao@adsc.com.sg, xhzhong@ntu.edu.sg

dl-jones@illinois.edu, aseschng@ntu.edu.sg, hli@i2r.a-star.edu.sg

## ABSTRACT

This paper presents a learning-based approach to the task of direction of arrival estimation (DOA) from microphone array input. Traditional signal processing methods such as the classic least square (LS) method rely on strong assumptions on signal models and accurate estimations of time delay of arrival (TDOA). They only work well in relatively clean conditions, but suffer from noise and reverberation distortions. In this paper, we propose a learning-based approach that can learn from a large amount of simulated noisy and reverberant microphone array inputs for robust DOA estimation. Specifically, we extract features from the generalised cross correlation (GCC) vectors and use a multilayer perceptron neural network to learn the nonlinear mapping from such features to the DOA. One advantage of the learning based method is that as more and more training data becomes available, the DOA estimation will become more and more accurate. Experimental results on simulated data show that the proposed learning based method produces much better results than the state-of-the-art LS method. The testing results on real data recorded in meeting rooms show improved root-mean-square error (RMSE) compared to the LS method.

**Index Terms**— microphone arrays, direction of arrival, least squares, machine learning, neural networks.

## 1. INTRODUCTION

The direction of arrival (DOA) estimation of a sound source using microphone arrays in noisy and reverberant environments is an important task in many applications such as distant automatic speech recognition [1, 2], teleconferencing [3], and automatic camera steering [4]. However, accurate DOA estimation can be very challenging when the received speech signals are significantly distorted due to background noises and room reverberations. An approach for the robust DOA estimation in such conditions is highly demanded.

Over the last few decades, a wide range of signal processing approaches are developed for the DOA estimation in noisy and reverberant environments. These approaches can be generally divided into following categories: *i*) the subspace based approaches such as the multiple signal classification (MUSIC) [5, 6] and the estimation of signal parameters via rotational invariance techniques (ES-PRIT) [7], *ii*) the time delay of arrival (TDOA) approaches using the generalized cross correlation method [8] and the least squares (LS) method [9], *iii*) the signal synchronization based approaches such as the steered response power with phase transform (SRP-PHAT)

[10], the multichannel cross correlation (MCCC) [11], the averaged magnitude difference function (AMDF) and the averaged magnitude sum function (AMSF) [12], *iv*) the approaches based on the blind identification of impulse responses such as the adaptive eigenvalue decomposition (AED) algorithm [13] and the independent component analysis method [14], *v*) the  $l_1$ -norm penalty based sparse signal representation approaches [15, 16], and *vi*) the model based approaches such as the maximum likelihood method (MLM) [17] and the precedent echo effect modelling method [18]. In practice, the above approaches may face either one or a combination of following problems: high computational cost, non-realistic assumptions on signal/noise models, and unreliable performance in real environments [19].

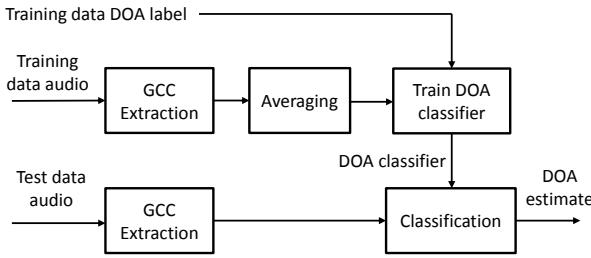
In this paper, we propose a learning-based approach to address the DOA estimation in noisy and reverberant environments. Unlike the existing methods which mainly rely on the array geometry and a short signal observation to estimate the DOA (e.g., the LS method), the proposed approach directly learns the nonlinear relationship between the received signals and the DOA from a large amount of training data synthesised for many noisy and reverberant environments. We will show that if a testing environment has the noise and reverberant conditions similar to one of the trained environments, the learning-based method can achieve very accurate DOA estimation based on its “memory” learnt from the training process. As such, the learning-based method is potentially more robust in challenging environments, i.e., the environments with low signal to noise ratio (SNR) and heavy reverberation. Furthermore, as more synthetic training data are available, a real environment is more likely to be matched and DOA estimation performance can be more reliable.

The rest of this paper is organised as follows. Section 2 describes the proposed approach that treats the DOA estimation problem as a pattern classification problem. Section 3 presents the performance study of the proposed approach from both simulations and the real speech recording experiments. Finally, some conclusions are drawn in section 4.

## 2. LEARNING-BASED DOA ESTIMATION APPROACH

We are interested in generating a system that can estimate DOA of a speech source when the microphone array geometry is known, e.g. an 8-channel circular array is used in our study. It is known that there is a nonlinear relationship between the TDOA of each microphone pair and the DOA. Classic LS-based DOA estimation meth-

**Fig. 1.** System diagram of the proposed learning-based approach for DOA estimation.



ods make use of this relationship and estimate the DOA using an iterative method. While the LS method works well in clean conditions, its performance degrades significantly when the background noise and/or reverberation are strong. Reliable DOA estimation in adverse environments is challenging using only the observed microphone signals. A possible solution to this problem is to rely on relevant prior knowledge about the problem. In this study, we would like to investigate how to use a big set of training data to achieve reliably DOA estimation in adverse environments. The idea is to learn a mapping from features extracted from the microphone array inputs to the DOA using a big set of training data.

We formulate the DOA estimation as a 360-class pattern classification problem<sup>1</sup>, where the microphone array inputs are classified to a DOA class from 0 degrees to 359 degrees. The classification system comprises of training and test phases, as illustrated in Fig. 1. In the training phase, a DOA classifier, a multilayer perceptron (MLP) neural network [20] in our case, is trained from a training data set, which contains pairs of array inputs and their DOA labels. For each audio segment (e.g. 0.2s long), a fixed dimension feature vector is extracted from GCC. In the test phase, the DOA classifier is used to generate the posterior probabilities of all the 360 DOA classes given a feature vector, and the class with the highest probability is selected as the estimated class. In the following sections, the details of the feature extraction, DOA classifier training, and DOA estimation will be described in details.

## 2.1. GCC as feature representation for DOA estimation

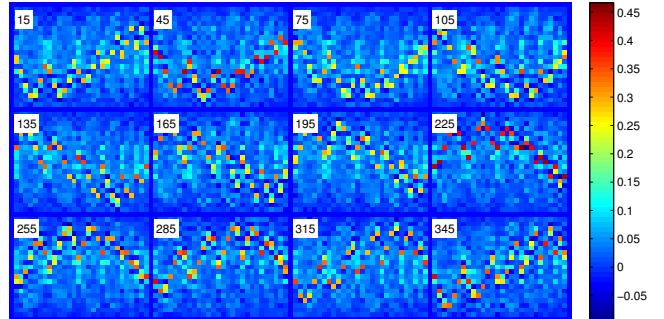
The first task is to choose a representation that contains sufficient information for the DOA estimation task. In the LS method, the TDOAs of microphone pairs are estimated from their GCC vectors. If TDOAs are estimated correctly, they are sufficient for DOA estimation. However, the TDOA estimation is often unreliable in low SNR and high reverberation conditions. Therefore, they are not suitable to be used as features for the DOA classifier. Compared to the TDOAs, the GCC patterns are more reliable and contain all the information required for the DOA estimation. Therefore, the GCC is chosen as feature representation of the MLP model in our study.

We now use a concrete example to describe the extraction of GCC vectors. Suppose we use an 8-channel circular array with a diameter of 20 cm<sup>2</sup>. There are totally  $C_2^8 = 28$  pairs of microphones

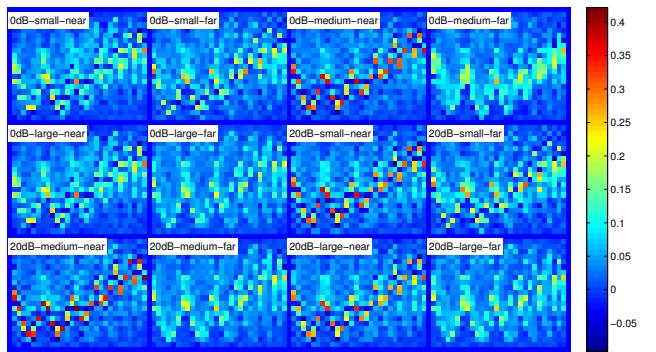
<sup>1</sup>It is not compulsory to use 360 classes. Other number of classes, e.g. 180 or 720, may also be used. We also tried to formulate the DOA estimation as a regression problem. However, it works worse than the classification formulation.

<sup>2</sup>Our approach can also be extended to other array configurations.

**Fig. 2.** GCC patterns for different DOA degrees in clean condition.



**Fig. 3.** GCC patterns for DOA=13 degrees in different SNR levels, room sizes, and distances between sound source to the array.



in the array, from each of them a GCC vector is computed for every 0.1s using the generalized cross correlation method with phase transform (GCC-PHAT) [8] for its good robustness to reverberation. The maximum possible time delay between any 2 microphones is  $\tau = 0.2/340 = 0.5882ms$  where the sound speed is assumed to be 340m/s. Suppose we are using a sampling rate of 16kHz, then the maximum delay in samples is  $n = 16000\tau \approx 21$ . Hence, for each microphone pair, only the 21 correlation coefficients near the center<sup>3</sup> contains useful information for DOA estimation. If we arrange all the GCC vectors from the 28 microphone pairs to a  $21 \times 28$  matrix, interesting patterns can be observed as shown in Fig. 2. In the figure, different DOAs correspond to different GCC patterns. Therefore, the DOAs can be potentially inferred from the GCC patterns.

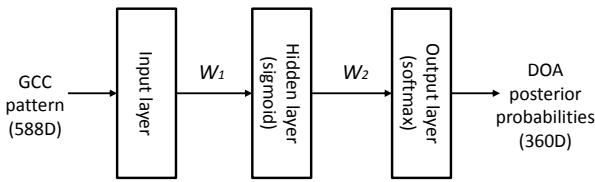
The GCC patterns are not only determined by the DOA, but also distorted by other nuisance factors, such as noise and reverberation. In Fig. 3, we show 12 GCC patterns computed from the same DOA in different room conditions. It can be observed that the 12 GCC patterns are very different in different SNR levels, rooms, and distances conditions, although some similarities of the GCC patterns can still be perceived. Hence, for robust DOA estimation, the DOA classifier needs to generalise well to unseen test conditions, such as unseen room configurations, noise types, and reverberation characteristics.

## 2.2. DOA estimation as a classification problem

Given the GCC patterns as input, the DOA classifier is used to generate the posterior probability of all the 360 DOA angles. Let  $\mathbf{o}_t$  de-

<sup>3</sup>The centre is the correlation coefficient corresponding to TDOA=0.

**Fig. 4.** The architecture of the MLP-based DOA classifier.



notes the  $588 \times 1$  feature vector obtained by converting the  $21 \times 28$  GCC pattern into vector format and  $t$  is the frame index. The posterior probability of a DOA angle is  $p(\theta_t | \mathbf{o}_t)$  for  $\theta_t = 0, \dots, 359$  degrees. To predict the posterior probabilities, we choose to use MLP as it can handle high dimensional inputs without making any assumption to the input data distribution. The structure of the MLP-based DOA classifier is shown in Fig. 4. The input layer of the MLP has the same number of nodes as the input feature dimension, i.e. 588 in our case. There is only one sigmoid hidden layer in the illustrated network although more hidden layers are possible. The weight matrix  $W_1$  between input features and hidden layer has a dimension of  $H \times 588$  where  $H$  is the number of hidden nodes. The activations of the hidden nodes are converted to DOA class posterior probabilities by using a linear transformation with weights  $W_2$  (size  $360 \times H$ ) and the softmax activation function. Mathematically, the posterior probabilities are nonlinear functions of the input features:

$$\mathbf{a}_t = f(W_1 \mathbf{o}_t + b_1) \quad (1)$$

$$\mathbf{z}_t = W_2 \mathbf{a}_t + b_2 \quad (2)$$

$$p(\theta_t = i | \mathbf{o}_t) = \frac{\exp(\mathbf{z}_t(i))}{\sum_{j=1}^C \exp(\mathbf{z}_t(j))}, \quad i \in [0, C - 1] \quad (3)$$

where  $f(x) = 1/(1 + e^{-x})$  is the sigmoid function and applied to the elements of its input individually, and softmax function is used in equation (3).  $C = 360$  is the number of classes,  $b_1$  and  $b_2$  are biases for the hidden layer and output layer, respectively.

The MLP is trained from a training data set  $\{\{\mathbf{o}_t, \theta_t\} | t = 1, \dots, T\}$  and  $T$  is the total number of samples (frames) in the training set. During the preparation of the training data, the DOA of a sound source is made to remain constant for several seconds. Multiple GCC vectors computed by using 0.1s frame shift can be generated. All the GCC vectors use the same DOA label. Some of these vectors correspond to silence portion of the recording and do not contain the useful patterns as shown in Fig. 2. Therefore, it is not suitable to use GCC vectors from silence portions to train the MLP. As we are using clean speech signals to generate simulated training data, accurate voice activity detection (VAD) can be obtained and only GCC vectors from speech segments are used for training. The stochastic gradient descent (SGD) algorithm [20] can be used to train the MLP iteratively.

### 2.3. Robust DOA classification

The DOA angle for a test recording needs to be estimated. In this study, we constrain the test recording to be several seconds long. Considering that the GCC pattern of a single test frame is noisy and may come from non-speech frames, we proposed to use a weighted

**Table 1.** Configurations used for generating training and test data. All rooms are 3m high. Distances are between array and source.

Simulated Training Data	
Speech	7861 sentences from WSJCAM0 training set
Room size (m)	small (7x5), medium (12x10), large (17x15)
Distance (m)	near (1) and far (2, 4, 6.5 for small, medium, large)
T60 (s)	0.1s to 1.0s with 0.1s step
SNR (dB)	Uniformly sampled from 0 to 20dB

Simulated Test Data	
Speech	538 sentences from WSJCAM0 et1 test set
Room size(m)	small (6x4), medium (10x8), large (14x12)
Distance (m)	near (1) and far (1.5, 3, 5 for small, medium, large)
T60 (s)	0.3s for small, 0.6s for medium, 0.9 for large room
SNR (dB)	3 categories: 0dB, 10dB, and 20dB

sum of the GCC patterns from all the frames of the test recording:

$$\mathbf{o} = \sum_{m=1}^L w_m \mathbf{o}_m \quad (4)$$

$$w_m = \frac{\sum_{d=1}^D |o_m(d)|^\alpha}{\sum_{m=1}^L \sum_{d=1}^D |o_m(d)|^\alpha} \quad (5)$$

where  $L$  is the number of frames in the test recording and  $w_m$  is the weight of frame  $m$ .  $D = 588$  is the dimension of GCC vectors,  $|\cdot|$  takes absolute value, and  $\alpha$  is a control parameter. If  $\alpha = 0$ , the mean of the GCC vectors is obtained. A large  $\alpha$  should be used to reduce the contribution of silence frames. It is from the observation that the GCC vectors from speech frames usually contain strong correlation coefficients near 1, while GCC vectors from silent frames contain weak correlation coefficients near 0.

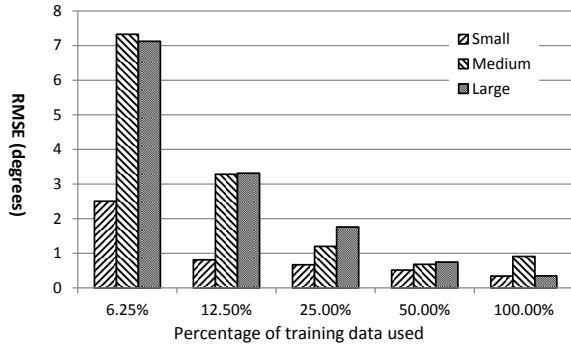
To improve the robustness of the DOA classifier, we also applied two feature normalisation methods to the mean GCC vector. The first normalisation is the histogram equalisation (HEQ) and widely used in image processing [21] and noise robust speech recognition [22, 23, 24]. The HEQ is used to normalise the histogram of the test GCC pattern to the average histogram of training GCC patterns to reduce the mismatch between test and training GCC patterns. We also scale the HEQ-processed GCC pattern such that its max value is 1. This is from the fact that the theoretical maximum correlation coefficients in GCC vector is 1, but the actual maximum value of a GCC vector may be quite low, e.g. 0.2, due to noise and reverberation.

## 3. EXPERIMENTS

### 3.1. Experimental Settings

The proposed learning based approach is evaluated and compared to the classic LS method [9] for DOA estimation on both simulated and real data. The 8-channel circular array with a diameter of 20cm is used. The simulated data is synthesized by convolving clean speech signals with the room impulse responses (RIRs) measured from the array based on the image method [25], and the additive noises are added. Variations of the simulated data were made with different room sizes, reflection rates, source to array distances, and SNR levels. The real test data was recorded in a small (6x4m) and a large (10x7m) meeting rooms using a real circular array. The source to array distances are 1.5 m and 3 m in the small meeting room and 6 m in the large meeting room.

**Fig. 5.** Average RMSE of the DOA classifier on the simulated test data of 3 rooms using different amount of training data. 100% refers to the case where all 47,166 simulated training recordings are used.



**Table 2.** Performance of LS and MLP methods on Simulated Data.

Room	Method	SNR=0dB		SNR=10dB		SNR=20dB	
		Near	Far	Near	Far	Near	Far
RMSE							
Small	LS	18.64	27.62	10.97	8.33	7.07	6.19
	MLP	0.45	1.26	0.11	0.50	0.00	0.44
Medium	LS	15.02	54.70	9.40	21.74	7.18	9.39
	MLP	0.20	6.17	0.00	0.73	0.00	0.61
Large	LS	14.80	54.98	6.02	17.21	1.08	13.37
	MLP	0.12	7.33	0.05	0.83	0.00	0.75
Mean absolute error (MAE)							
Small	LS	5.79	10.98	1.72	1.52	1.39	1.33
	MLP	0.14	0.53	0.01	0.24	0.00	0.19
Medium	LS	4.97	31.36	1.81	7.05	1.47	2.99
	MLP	0.04	1.53	0.00	0.44	0.00	0.34
Large	LS	4.69	31.29	1.48	7.21	0.74	4.26
	MLP	0.01	1.36	0.00	0.54	0.00	0.50

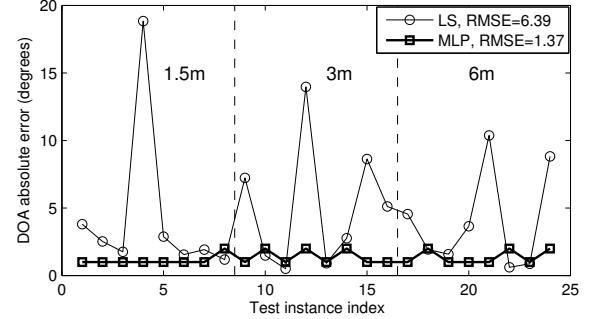
The simulated data are divided into 2 sets with different settings (see Table 1) for the training of the DOA classifier and performance testing. In the synthesis of the training data, we convolved 7861 clean sentences from the WSJCAM0 [26] training set with the simulated RIRs of the array. Then the additive noises taken from the Reverb Challenge 2014 corpus [27] were added to the training data with the SNR levels randomly chosen from 0dB to 20dB for each sentence. Each of the 7861 clean sentences was used 6 times with randomly selected RIR, noise signals, and SNR levels. Totally 47,166 training sentences were synthesised for the 360 angles. The MLP model has one hidden layer with 512 hidden nodes.

Similar to the training data, the simulated test data is generated by convolving 538 clean utterances from WSJCAM0 test set with an average sentence length of 6.9s. For each room configuration, 360 test utterances are synthesised for 360 DOA angles. During testing,  $\alpha$  was set to 4 in (4). HEQ and maximum value normalisation are applied in sequence on the averaged GCC patterns. The DOA estimation performance is evaluated by two measures: the mean of absolute error (MAE) and the root mean square error (RMSE).

### 3.2. Results

In Fig. 5, we show the average RMSEs of the MLP-based approach for the DOA estimation using different amounts of training data. It demonstrates that as more training data is used, the RMSE of the

**Fig. 6.** Comparison of LS and MLP performance on real recordings. The 24 recordings were taken with distances of 1.5m, 3m, and 6m. For each distance, 8 DOA were used, from 0 to 315 degrees with a step size of 45 degrees.



MLP-based approach decreases. This verified our claim that the performance of the MLP-based approach is keeping improved when relevant training data size becomes larger. The traditional methods such as LS does not make use of training data, hence its performance cannot be improved from the available data.

We compared the DOA estimation performance of the MLP method with the LS method on the simulated data in Table 2. We used the full percentage of the training data for the MLP method. It is observed that for all the test conditions the MLP method outperforms the LS method significantly, especially on the 0dB data where the performance of the LS method degrades dramatically. The results also demonstrate the robustness of the MLP method when the test data can be matched by the training data. In fact, we may take this advantage for the real applications in a known environment. A training data set that tries to match the real environment may be synthesized. We will study this further in our future work.

In Fig. 6, we show the DOA estimation results of the MLP and LS methods on the real recordings. Without doing any matching from the training data to the real test environment, the MLP method obtained much smaller absolute errors than the LS method. The RMSE of the MLP method is 1.37, which is also much smaller than that of the LS method. We will study the possibility of matching a training data to the given test environment in our future work.

## 4. CONCLUSIONS

In this paper, we proposed a learning-based approach using MLP to the DOA estimation problem in challenging environments. A classification problem for the DOA estimation was presented. The MLP input features and the robust DOA classification were discussed. We have demonstrated that a MLP model which is trained on simulated data can learn a regularity which maps the GCC patterns to the DOAs. Experimental results on the simulated test data showed that the learnt mapping is robust to high level noises and strong reverberations. We also demonstrated the performance of the MLP method on the real recording data and superior performance over the LS method was obtained. We believe that the true potential of the proposed learning-based approach could be beyond the results shown in this study. Using larger amount of training data, or trying to match the training data to a test environment, or using multiple hidden layers (i.e., deep neural networks) may further improve the performance.

## 5. REFERENCES

- [1] M. Woelfel and J. McDonough, *Distant Speech Recognition*, Wiley, 2009.
- [2] X. Xiao, S. Zhao, D. H. H. Nguyen, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, “The NTU-ADSC systems for reverberation challenge 2014,” in *Proceedings of the Reverberation Challenge Workshop*, 2014.
- [3] S. Zhao, S. Ahmed, Y. Liang, K. Rupnow, D. Chen, and D. L. Jones, “A real-time 3D sound localization system with miniature microphone array for virtual reality,” in *Proceedings of the 7th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, 2012.
- [4] S. Zhao, E. S. Chng, N. T. Hieu, and H. Li, “A robust real-time sound source localization system for olivia robot,” in *Proceeding of the APSIPA Annual Summit and Conference*, 2010.
- [5] R. O. Schmidt, “Multiple emitter location and signal parameter estimation,” *IEEE Trans. Antennas Propagat.*, vol. AP-34, pp. 279–280, March 1986.
- [6] S. Zhao, T. Saluev, and D. L. Jones, “Underdetermined direction of arrival estimation using acoustic vector sensor,” *Signal Processing*, vol. 100, pp. 160–168, 2014.
- [7] R. Roy and T. Kailath, “Esprit-estimation of signal parameters via rotational invariance techniques,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 7, pp. 984–995, July 1989.
- [8] C. H. Knapp and G. C. Carter, “The generalized correlation method for estimation of time delay,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, August 1976.
- [9] Y. Huang, J. Benesty, G.W. Elko, and R.M. Mersereau, “Real-time passive source localization: A practical linear-correction least-squares approach,” *IEEE Trans. Speech and Audio process.*, vol. 9, November 2001.
- [10] J. H. Dibiase, “A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays,” in *Ph.D. dissertation*. Brown Univ., Providence, RI, 2001.
- [11] J. Benesty, J. Chen, and Y. Huang, “Time delay estimation via linear interpolation and cross-correlation,” *IEEE Trans. Speech and Audio Process*, vol. 12, no. 5, September 2004.
- [12] M. Souden, J. Benesty, and S. Affes, “Broadband source localization from an eigenanalysis perspective,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 18, no. 6, August 2010.
- [13] J. Benesty, “Adaptive eigenvalue decomposition algorithm for passive acoustic source localization,” *J. Acoust. Soc. Amer.*, vol. 107, pp. 384–391, January 1989.
- [14] A. Lombard, Y. Zheng, H. Buchner, and W. Kellermann, “Tdoa estimation for multiple sound sources in noisy and reverberant environments using broadband independent component analysis,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 19, no. 6, pp. 1490–1503, August 2011.
- [15] D. Malioutov, M. Cetin, and A.S. Willsky, “A sparse signal reconstruction perspective for source localization with sensor arrays,” *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 3010–3022, August 2005.
- [16] B. Wang, J. Liu, and X. Sun, “Mixed sources localization based on sparse signal reconstruction,” *IEEE Signal Process. Lett.*, vol. 19, no. 8, pp. 487–490, August 2012.
- [17] P. Stoica and K. Sharman, “Maximum likelihood methods for direction-of-arrival estimation,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 7, pp. 1132–1143, July 1990.
- [18] J. Huang, N. Ohnishi, and N. Sugie, “Sound localization in reverberant environment based on the model of the precedence effect,” *IEEE Trans. Instrum. Meas.*, vol. 46, no. 4, pp. 842–846, August 1997.
- [19] N. T. N. Tho, S. Zhao, and D. L. Jones, “Robust doa estimation of multiple speech sources,” in *Proceeding of ICASSP*. IEEE, 2014, pp. 2287–2291.
- [20] S. S. Haykin, *Neural networks and learning machines*, vol. 3, Pearson Education Upper Saddle River, 2009.
- [21] Y.-T. Kim, “Contrast enhancement using brightness preserving bi-histogram equalization,” *IEEE Transactions on Consumer Electronics*, vol. 43, no. 1, pp. 1–8, 1997.
- [22] A. De La Torre, A. M. Peinado, J. C. Segura, J. L. Pérez-Córdoba, M. C. Benítez, and A. J. Rubio, “Histogram equalization of speech representation for robust speech recognition,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 355–366, 2005.
- [23] X. Xiao, J. Li, E. S. Chng, and H. Li, “Maximum likelihood adaptation of histogram equalization with constraint for robust speech recognition,” in *Proceeding of ICASSP*. IEEE, 2011, pp. 5480–5483.
- [24] X. Xiao, E. S. Chng, and H. Li, “Attribute-based histogram equalization (heq) and its adaptation for robust speech recognition..,” in *Proceeding of INTERSPEECH*, 2013, pp. 876–880.
- [25] J.B. Allen and D.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, April 1979.
- [26] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, “WSJ-CAM0: a british english speech corpus for large vocabulary continuous speech recognition,” in *Proceeding of ICASSP*, 1995, pp. 81–84.
- [27] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas, S. Gannot, and B. Raj, “The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech,” in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA-13)*, 2013.