INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Feature Fusion by Attention Networks for Robust DOA Estimation

Rongliang Liu[1], *Nengheng Zheng*[1,2], *and Xi Chen*[1]

[1]The Guangdong Key Laboratory of Intelligent Information Processing, College of Electronic and Information Engineering, Shenzhen University, China
[2]Peng Cheng Laboratory, Shenzhen, China
1810262039@email.szu.edu.cn, nhzheng@szu.edu.cn, 1900432053@email.szu.edu.cn

Abstract

Direction of arrival (DOA) estimation is a key front-end technology for many speech-based intelligent systems. Deep neural networks-based DOA systems have recently demonstrated better performances than conventional ones. However, most of the existing networks use only one specific acoustical feature as input, limiting their noise-robustness. This paper proposes an attention-based feature fusion approach for DOA estimation. Two classical DOA estimation approaches, i.e., the least mean square-based adaptive filtering and the generalized cross-correlation, are adopted, and the respective features are served as input to the networks. Network with attention mechanism is built to learn the optimal weighting scheme, which can take advantage of the two features' complementary contributions in DOA estimation. Simulation and real test results show that the proposed method could use the complementary DOA information in different features and improve estimation accuracy under acoustic conditions with both noise and reverberation.

Index Terms: DOA estimation, attention-mechanism, feature fusion, GCC feature, LMS feature

1. Introduction

DOA estimation is a key component for intelligent human-machine interaction devices, and it has achieved good performance in applications such as teleconferencing [1], robotics [2] and beamforming [3]. However, accurate DOA estimation is still challenging in real-world environments with strong background noise and reverberation [4].

Conventional DOA estimations can be divided into three categories: *i*) Time-delay based, in which the arriving time difference between microphones can be computed with various methods, e.g., the generalized cross-correlation (GCC) [5,6], the least mean square-based adaptive filtering [7], etc. *ii*) Beamforming-based, in which the direction of the strongest signal is detected with beamforming, e.g., steered response power with phase transform (SRP-PHAT) [8]. *iii*) High-resolution spectral estimation based, which calculates the covariance matrix of signals collected by the microphones to get the spatial spectrum, e.g., the multiple signal classification (MUSIC) [9] and the estimation of signal parameters via rotational invariance technique (ESPRIT) [10]. These methods, however, generally suffer from some inevitable problems, e.g., unreasonable assumption on signal/noise models [11]; high computational costs, especially for the SRP-PHAT, MUSIC and ESPRIT; and dramatically degraded performance in strong noise and reverberation environments [12].

Recently, neural networks-based deep learning frameworks have been adopted for DOA estimation. For example, a multilayer perceptron (MLP) network, fed with GCC features, was built to learn the mapping relationship between features and DOA in noisy and reverberant environments [13]. A similar framework was adopted for multi-speaker DOA estimation for a robot [14]. In [15], the speaker position coordinates were estimated using a deep neural network (DNN) with GCC features. A convolutional neural network (CNN) with phase features was introduced for DOA estimation in [16]. Compared to traditional algorithms, deep learning methods achieved certain robustness against noise and reverberation. However, most of them used only one kind of acoustic feature. Complementary contributions from different features have not been fully investigated for DOA estimation.

This paper proposes an attention neural network to take advantage of two complementary features for DOA estimation. A feature attention module (FAM) is first built to generate weights for two acoustical features, i.e., the GCC feature and the least mean square adaptive filtering (LMS) features. Then a feature fusion module (FFM) is constructed to train a classifier with the weighted feature combination. Finally, a set of simulations and real-world experiments are conducted for performance evaluation.

2. System description

2.1. DOA estimation as a classification problem

Unlike the conventional DOA estimations, which compute the azimuth, w.r.t. the microphone array, of the coming sounds, the deep learning-based methods formulate the estimation as a *K*-class classification problem. For example, for a uniformly linear array (ULA) with DOA from 0° to 180°, there is K = 19 DOA classes given a resolution of 10°. Let $\Theta=\{\theta_k\colon k=1,2,\cdots,K\}$ be the DOA set containing K classes, DOA estimation for an input sound x can be solved by searching for θ_k with the largest posterior probability, i.e.,

$$\hat{k}=\underset{k}{\operatorname{argmax}} P(\theta_k|x) \tag{1}$$

where $\hat{k}$ denotes the index for the most probable element of the set. The resolution and the K can be arbitrarily determined depending on the tasks.

In this study, the classification is a supervised learning process. The DOA classifier is trained using a large amount of simulation data, with combined feature maps and corresponding DOA labels. In the test, given a combined feature map, the classification system outputs the posterior probability for each of the K DOA classes, and the label with the largest probability is selected as the estimated DOA.

http://dx.doi.org/10.21437/Interspeech.2021-1051

2.2. Proposed model architecture

An attention-based feature fusion network is constructed for the DOA classification. As shown in Fig. 1, the system contains a feature attention module (FAM) and a feature fusion module (FFM). The FAM consists of a global average pooling layer, two convolutional layers, and a rectified linear units (ReLU) [17] activation function between the two convolutional layers. Each convolutional layer has 16 local filters of size 1 × 1. The input to FAM is a dual-channel feature map consisting of two different features. The FAM is trained to generate weighting parameters for each input feature under various environmental conditions. To do so, a channel descriptor is first applied to each of the input features to compute (by global average pooling) a descriptor f_c for the feature, i.e.,

$$f_c = \boldsymbol{F}(x_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i,j) \tag{2}$$

where $x_c(i,j)$ denotes the feature value of channel c at position (i,j), $\boldsymbol{F}$ is the global pooling function, H and W are the length (feature-length of one pair of microphones) and width (all microphone pairs) of the input feature map. That is, the input feature map is now converted into low-dimensional channel descriptors, which are then passed through two convolution layers, respectively activated by ReLU and Sigmoid function, to get the weight of each channel, i.e.,

$$w_c = \boldsymbol{\sigma}(Conv(\boldsymbol{\rho}(Conv(f_c)))) \tag{3}$$

where w_c is the weight for channel c, $\boldsymbol{\sigma}$ is the Sigmoid function, and $\boldsymbol{\rho}$ is the ReLU activation function. Finally, flatten the input feature x_c to one-dimensional vector x_c^1, then multiply x_c^1 by the corresponding channel attention weight to get the enhanced feature x_c':

$$x_c' = w_c \otimes x_c^1 \tag{4}$$

The FFM consists of two LSTM layers and two fully connected layers (FCL). This module uses LSTM to learn the DOA information between different frames to enhance the performance of the system. The number of nodes is 128 for each LSTM layer, 512 for FCL and 19 for the output layer. A Batch normalization layer [18] with the tanh activation function is implemented after the first FCL (not illustrated in Fig. 1). Cross-entropy [19] and Adam [20] optimizer are used to train the model. A dropout procedure [21] with a rate of 0.5 and early-stop with 15 epochs are used to avoid overfitting.

2.3. GCC and LMS feature extraction

Time difference of arrival (TDOA) based methods are the most popular ones for conventional DOA estimation. In this study, two classic TDOA methods, the generalized cross-correlation with phase transform (GCC-PHAT) and the least mean square adaptive filtering, are adopted to extract the features.

GCC features are widely used as they contain useful information for DOA estimation. In this work, we use a 6-channel ULA with an inter-microphone distance of 4cm. The GCC-PHAT between any two microphones, denoted the ith and the jth, is formulated as

$$GCC_{ij}(\tau) = \sum_{\omega} \boldsymbol{R}\left(\frac{X_i(\omega)X_j^*(\omega)}{\left|X_i(\omega)X_j^*(\omega)\right|} e^{j\omega\tau}\right) \tag{5}$$

where τ is the delay in the discrete domain, * denotes the complex conjugation and $\boldsymbol{R}$ means taking the real part of a complex number. There are $C_6^2 = 15$ pairs of GCC features to be computed. Given the sound speed of 340m/s and the maximum inter-microphone distance of 20cm, the maximum time delay $\tau \approx$ 0.588ms, which (at the sampling rate of 16kHz) results in a maximum delay of 21 samples. Therefore, the extracted feature map is a 15 × 21 matrix for each time frame.

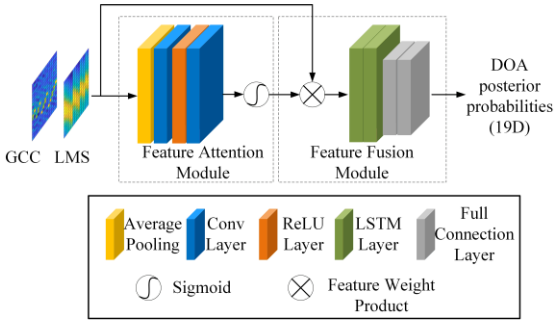


Figure 1: *Architecture for the attention-based DOA classification system.*

Adaptive filtering with the least mean square criterion was adopted for DOA estimation in the early years. Let signals received at two microphones, $x_1(n)$ and $x_2(n)$, be the input to and the output from a filter $h(n)$, an optimal output $\hat{x}_2(n)$ from the filter can be obtained by adaptively updating $h(n)$ such that the mean square error between the filter output and the target is minimized, i.e.,

$$\hat{x}_2(n) = x_1(n) * h_{opt}(n) \tag{6}$$

where

$$h_{opt}(n) = \underset{h(n)}{\operatorname{argmin}} \boldsymbol{E}\{x_2(n), x_1(n) * h(n)\} \tag{7}$$

$h_{opt}(n)$ can be used for DOA estimation. In this study, similar to the GCC feature map, a 15 × 21 matrix can be constructed at each time frame to be an LMS feature map.

Figure 2 shows the two feature maps computed from two microphone-received signals with DOA of 60° and SNR of 0 dB and 15dB. As shown, at low SNR (0dB), the LMS feature has much more obvious peaks than GCC, which tells that the LMS contains more useful information for DOA at low SNRs. On the other hand, although both features show peaks at high SNR (15dB), the GCC feature has a more prominent single peak than the LMS features, which tells that the GCC could be more useful in DOA estimation at high SNRs. The networks' attention-mechanism abovementioned are expected to take advantage of such complementary contributions for DOA estimation, especially in noisy conditions.

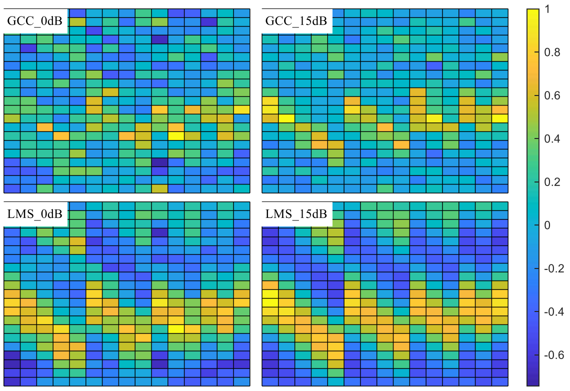


Figure 2: *Examples of GCC (top) and LMS (bottom) feature maps with DOA=* 60°*, SNR=0dB and 15dB, and the source-to-array distance of 1m.*

3. Simulation experiments

3.1. Experimental setup

DOA estimation experiments were conducted in a simulated environment. The rooms simulated were of two different sizes, i.e., R1: $6\text{m} \times 8\text{m} \times 3\text{m}$ and R2: $10\text{m} \times 9\text{m} \times 3\text{m}$. The adopted ULA was a six-microphone array with a 4cm inter-microphone distance. It was placed at the center of 1m from the wall, and its height was fixed at 1.5m. The horizontal distance between the sound source and ULA was changed in different room sizes, and the height of the source was 1.7m. The horizontal azimuth of the source ranged from 0° to 180° in an interval of 10° .

The Image method [22] was adopted to generate the room impulse response (RIR) at different conditions such as room size, source-to-array distance, and reverberation time (T60). Clean speech from the TIMIT database [23] was used as the source speech. The microphone-received speech was simulated by

$$x(n) = s(n) * h_r(n) + e(n) \tag{8}$$

where $x(n)$ denotes simulated speech, $s(n)$ is the clean speech from the source, $e(n)$ is the noise selected from the Noise-92 database [24], $h_r(n)$ is the RIR.

The simulated data were divided into training and test sets. In synthesizing the training data, the source-to-array distances were divided into two types, the near (1m) and the far (2m). The reverberation time (T60) was set to be from 0.2 to 0.8s with a 0.2s step. White and babble noises were added with randomly chosen SNRs between 0 and 20dB. A voice activity detector (VAD) [25] was implemented to exclude the silent segments. In total, the training data contains approximately 0.56 million frames for the 19 DOA classes. A learning rate of $1e-3$ and a batch size of 512 frames were set for training.

The test data were generated in a similar way to the training ones except that: room sizes were different, i.e., R1: 6m × 6m and R2:8m × 8m; the near distance was 1m for both rooms and the far distance was 2.5m for R1 and 3.5m for R2; the T60 of R1 and R2 was 0.3s and 0.7s, respectively. Furthermore, to evaluate the system's generalization ability, white, babble, pink and f16 noises were used.

3.2. Baseline system

The baseline system is a state-of-the-art DNN-based [13] DOA estimation using GCC features, denoted as DNN-GCC. The GCC feature extraction is the same as described in Sec. 2.3 with a dimension of 1×315. The baseline network consists of two hidden layers, 1024 nodes in each layer, and an output layer containing 19 nodes. The dropout rate of 0.5 and tanh activation is adopted for hidden layer training. In addition, a Batch normalization layer was applied after each hidden layer.

3.3. Results and discussion

The DOA estimation performance of different systems is evaluated by the frame-level accuracy, i.e.,

$$A_c = \frac{N_c}{N_t} \tag{9}$$

where N_t denotes the total number of test frames and N_c denotes the number of correctly estimated frames.

Table 1: *DOA estimation performance (in %) under different environmental conditions with two noise types.*

Room	Method	SNR=0dB		SNR=10dB		SNR=20dB	
		near	far	near	far	near	far
seen noisy type							
R1	DNN-GCC	51.09	29.95	67.72	42.57	80.05	48.20
	proposed	**56.76**	**33.05**	**73.59**	**43.45**	**83.29**	**53.02**
R2	DNN-GCC	44.39	20.82	60.91	28.00	73.62	31.22
	proposed	**49.29**	**23.26**	**65.70**	**29.61**	**78.06**	**32.76**
unseen noisy type							
R1	DNN-GCC	42.47	24.61	64.66	39.80	81.48	48.74
	proposed	**45.71**	**26.66**	**69.31**	**42.38**	**85.06**	**52.18**
R2	DNN-GCC	38.14	18.80	59.98	27.74	75.34	31.86
	proposed	**41.98**	**19.64**	**64.98**	**28.39**	**79.56**	**33.25**

Table 1 gives the estimation accuracy by two systems, the proposed and the baseline, under different environmental conditions. As shown, the proposed method outperforms the baseline under all conditions. Take R1 (T60=0.3s) with source-to-array distance (near) as an example, when 20dB seen noise presented, 83.29% accuracy can be achieved by the proposed system, compared to the 80.05% by DNN-GCC, a relative improvement of 4.05%; when SNR reduced to 0dB, the performances degrade dramatically for both systems, i.e., 56.76% v.s. 51.09%. Nevertheless, a relative improvement of 11.10% by the proposed method was witnessed. As for far source-to-array distance, performance degradation was also observed, i.e., 53.02% v.s. 48.2% at 20dB noise. Still, a superiority of 10% achieved by the proposed. Different reverberation also affects the estimation accuracy. As shown, in R2 (T60=0.7s), the accuracy of the proposed system is 78.06% compared to 73.62% by DNN-GCC. Again, a relative improvement of 6.03% is achieved. Similar phenomena can be observed when unseen noise is presented. The superiority of the proposed system might come from taking advantage of the complementary contribution from the two features, as shown in Fig. 2. in challenging environments.

An example of the two methods' performance is depicted in Fig. 3, which shows the frame-level output results for a speech sample at environmental conditions: white noise, SNR of 12dB, and room R2. The actual source DOA is 30°, 90° and 150°, which can be seen in the bottom subfigure. One can see that the output results from the proposed system (top) are much more focused on the original DOAs compared to the DNN-GCC (middle).

Figure 4 compares the distribution of the absolute estimation error, i.e., the difference between the true and the estimated DOAs, by the two systems. It is apparent that the means of errors at all conditions are much smaller by the proposed system, and also, whiskers of the proposed one are shorter than the baseline at most conditions. Comparing the angles on both sides (30° and 150°), the error range of the two methods is smaller at DOA of 90°, but the proposed one achieves a more accurate estimation with error very close to 0°. Besides, there are more outliers in far conditions, but on the whole, the proposed method achieves higher stability and robustness of DOA estimation.

Table 2 gives the number of parameters of the two methods. The proposed one has greater performance under the condition of fewer parameters compared with DNN-GCC.

Table 2: *Comparisons in terms of number of parameters*

Method	**# Param.**
DNN-GCC	3.44 M
Proposed	2.39 M

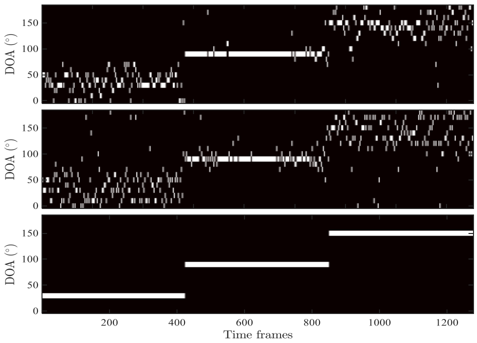


Figure 3: *Frame-level DOA estimation results for the proposed method (top), DNN-GCC (middle) and ground truth (bottom). From left to right, the true DOA is* 30°, 90° *and* 150°.

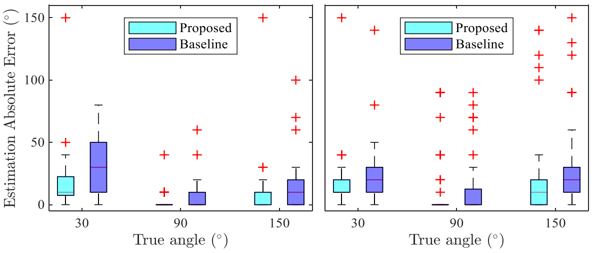


Figure 4: *Distribution of the absolute estimation errors for two systems in room R1 with 3 dB white noise. Left: near; Right: far. Box give the mean errors and whisker gives the upper and lower limits of the error value. The "+" indicates the outliers.*

4. Verification in real-world environments

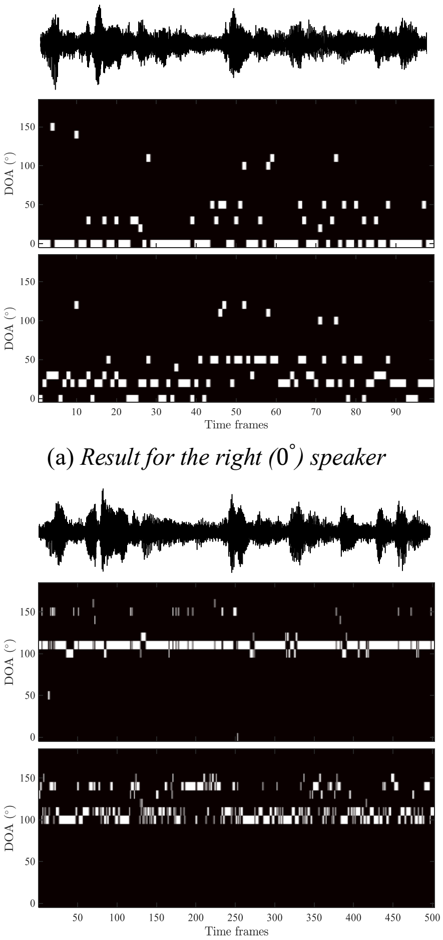
A practical DOA estimation system was set up to further verify the superiority of the proposed system over the baseline. To verify the generalization of the model, a six-microphone circular array with a 3.5cm radius was used. The training data were generated similarly to that in Section 3, but with different RIR and SNR configurations, as given in Table 3. The test data were collected in a real room with configuration as given in Table 4. Figure 5 gives two exemplar results at DOA of 0° and 110°. In Fig. 5(a), it can be seen that, noise caused some deviations in the estimation results of the two systems. But the proposed system locate the target speaker roughly at 0°, while the baseline system has more divergent results. As shown in Fig. 5(b), the proposed method forms an obvious white line at 110°, although some deviations exist. The baseline, however, output DOA estimated with a much more significant deviation. In general, compared with the baseline, the proposed system has significantly improved robustness in the real environment.

Table 3: *Configurations used for generating training data*

Items	Parameter
Room size (m)	3×3×2
Source-array distance (m)	near (1m) far (1.5m)
T60 (s)	0.2s to 0.8s with 0.2s step
SNR (dB)	Sampled from 0 to 20dB
DOA (°)	0° to 180° with 10° resolution

Table 4: *Configurations used for real test data*

Items	Parameter
Room size (m)	3.2×3.2×2.5
Speaker-array distance (m)	1.2m
Target speaker DOA (°)	right (0) and left (110)
Noise angle (°)	left (150)



(a) *Result for the right* (0°) *speaker*

(b) *Result for the front* (110°) *speaker*

Figure 5: *Frame-level DOA estimation results for the proposed system (middle), DNN-GCC (bottom). Top subfigure is the original signal waveform of one microphone.*

5. Conclusions

A neural network with an attention mechanism is proposed for robust DOA estimation. Two features implemented in the classical methods, i.e., the GCC and the LMS features, are adopted as input to the networks. The attention mechanism enables the network to take advantage of the complementary contributions from the two features. The superiority of the attention-based network over the baseline, DNN with GCC feature only, is verified by a set of experimental simulations and real testing.

6. Acknowledgments

This work is jointly supported by Guangdong Key Area R&D Project (No. 2018B030338001), National Natural Science Foundation of China (61771320). Nengheng Zheng is the corresponding author.

7. References

[1] Q. Yan, J. Chen, G. Ottoy and L. D. Strycker, "Robust AOA based acoustic source localization method with unreliable measurements," *Signal Processing*, vol. 152, pp. 13-21, Nov. 2018.

[2] S. Argentieri, P. Danes, and P. Soueres, "A survey on sound source localization in robotics: From binaural to array processing methods," *Computer Speech and Language*, vol. 34, no. 1, pp. 87–112, Nov. 2015.

[3] J. C. Chen, K. Yao and R. E. Hudson, "Source localization and beamforming," *IEEE Signal Processing Magazine*, vol. 19, no. 2, pp. 30-39, Mar. 2002.

[4] Q. Li, X. Zhang and H. Li, "Online Direction of Arrival Estimation Based on Deep Learning," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 2616-2620.

[5] C. Knapp and G. Carter, "The Generalized Correlation Method for Estimation of Time Delay," *IEEE transactions on acoustics, speech, and signal processing*, pp. 24(4): 320-327, Aug. 1976.

[6] Y. T. Huang, J. Benesty, G. W. Elko and R. M. Mersereati, "Real-time passive source localization: a practical linear-correction least-squares approach," *IEEE Trans. Speech Audio Process*, vol. 9, no. 8, pp. 943-956, Nov. 2001.

[7] D. Youn, N. Ahmed and G. Carter, "On using the LMS algorithm for time delay estimation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, no. 5, pp. 798-801, Oct. 1982.

[8] T. Long, J. Chen, G. Huang, J. Benesty and I. Cohen, "Acoustic Source Localization Based on Geometric Projection in Reverberant and Noisy Environments," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 143-155, Mar. 2019.

[9] R. Takeda and K. Komatani, "Performance comparison of MUSIC-based sound localization methods on small humanoid under low SNR conditions," *International Conference on Humanoid Robots (Humanoids)*, Seoul, 2015, pp. 859-865.

[10] R. Roy and T. Kailath, "Esprit-estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 73, no. 7, pp. 984-995, Jul. 1989.

[11] N. T. N. Tho, S. Zhao and D. L. Jones, "Robust DOA estimation of multiple speech sources," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, 2014, pp. 2287-2291.

[12] J. Benesty, J. Chen, Y. Huang, "Microphone Array Signal Processing," *Journal of the Acoustical Society of America*, vol. 125, no. 6, pp. 4097-4098, Apr. 2008.

[13] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. Cheng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2015, pp. 2814–2818.

[14] W. He, P. Motlicek and J. Odobez, "Deep Neural Networks for Multiple Speaker Detection and Localization," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, Brisbane, QLD, 2018, pp. 74-79.

[15] F. Vesperini, P. Vecchiotti, E. Principi, S. Squartini and F. Piazza, "A neural network based algorithm for speaker localization in a multi-room environment," in *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, Vietri sul Mare, 2016, pp. 1-6.

[16] S. Chakrabarty and E. A. P. Habets, "Broadband doa estimation using convolutional neural networks trained with noise signals," in 2017 *IEEE Workshop on Applications of Signal Processing to Au-dio and Acoustics (WASPAA)*, New Paltz, NY, 2017, pp. 136-140.

[17] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *2010 Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[18] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," in *PMLR*, Jun. 2015, pp. 448–456.

[19] J. Shore and R. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," *IEEE Transactions on information theory*, vol. 26, no. 1, pp. 26–37, Jan. 1980.

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Computer Science*, Dec. 2014.

[21] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, Jun. 2014.

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

[23] Z. Victor, S. Seneff and J. Glass, "Speech database development at MIT: Timit and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351-356, Aug. 1990.

[24] A. Varga, H. 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, Jul. 1993.

[25] J. Sohn, N. S. Kim and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1-3, Jan. 1999.