Learning Based DOA Estimation in Adverse Acoustic Environment using Co-prime Circular Microphone Array

Raj Gohil*, Aditya Raikar†, Gyanajyoti Routray* and Rajesh M. Hegde*
* Indian Institute of Technology , Kanpur , India
E-mail: {rgohil, groutray, rhegde}@iitk.ac.in
† TCS Research and Innovation – Mumbai , India
E-mail: aditya.raikar@tcs.com

Abstract—**The direction of arrival (DOA) estimation is a well-known research problem. It is conditional to different microphone array geometry and acoustic room conditions. It also becomes more challenging in the presence of noise and reverberation. Many traditional signal processing approaches such as least square (LS) based rely on time difference of arrival estimation which is not robust to adverse acoustic conditions and hampers the DOA estimation. This problem can be solved using learning-based algorithms, which uses a large amount of data simulated on similar acoustic conditions. Though much of the work in learning algorithms until now leverages augmentation techniques and deep neural network (DNN) architecture for achieving robustness in DOA estimation, very less attention is given to the feature representation. Robust feature representation can be achieved using certain geometry of microphone array. In this work, a framework comprising of a learning-based DOA estimation along with a circular co-prime microphone array(CCMA) arrangement is proposed. Experiment results show that a robust feature representation is indeed essential in estimating the DOA accurately and gives a significant improvement in terms of root mean squared error(RMSE) and mean-absolute error(MAE) scores when compared to other state-of-the-art DNN and signal processing approaches.**

I. Introduction

Microphone array-based DOA estimation is a well-known research problem and becomes more challenging in adverse acoustic conditions comprising reverberations and background noise. It has a number of applications, some of which are teleconferencing [1], camera steering [2], and automatic distant speech recognition [3], [4], which requires prior information of the source location. DOA estimates are affected due to distortions introduced by reverberations and background noise, and therefore there is a need for a robust DOA estimation algorithm.

Various signal processing algorithms over the past decade have been proposed to estimate DOA and can be categorized accordingly. First of such category is subspace based approaches which includes multiple signal classification (MUSIC) [5], [6]. Second category is based on time difference of arrival (TDOA) which uses Generalized Cross Correlation (GCC) method [7]–[10], and LS based method [11]. Third Category is based on generalization of the cross correlation algorithms such as steered response power with phase-transform (SRP-PHAT) [12], multi-channel cross correlation coefficient (MCCC) [13]. Fourth category comprises of probabilistic algorithms such as maximum likelihood(ML) method [14], Fifth Category includes methods based on histogram analysis [15], [16]. Most of the signal processing methods mentioned above get affected by acoustic distortions such as reverberation and background noise. DNN have been popular in the last decade because of the non-linearity it introduces to capture various complex patterns from the data. Due to a large amount of data and various simulation tools available, DNN has achieved the state of the art performance in various problem statements, including DOA estimation. Some of the noted work from last few years are referred in [17]–[19].

Different microphone geometries have been used for the DOA estimation like uniform linear array (ULA) [20], [21] and uniform circular array [22]. Recently co-prime circular microphone array was proposed, which utilizes the co-prime pair of arrays, where each sub-array is co-prime related [23]–[26]. The distance between the sensors in each sub-array is related to the number of sensors in the other sub-array. The co-prime array can resolve upto $\mathcal{O}(AB)$ sources (A and B co-prime number of elements in each sub-array) with $A+B-1$ sensors. This can be used when it is necessary to reduce the mutual coupling between the elements. The distance between the microphones in the co-prime microphone array is defined in such a way that the grating lobe problem was also minimized, which can eventually lead to more accurate DOA estimation as compared to a uniform circular array.

Though microphone geometry is not that important in learning-based algorithms [17], a robust feature representation in reverberate and noisy conditions is important to estimate DOA more accurately. The contribution of this paper is two folds; first, it analyses the feature representation in adverse conditions by studying feature extracted using a uniform circular microphone array(UCMA) and co-prime circular microphone array(CCMA). Second, the significance of array

[†] The second author was a part of this work when he was pursuing Masters degree at Indian Institute of Technology, Kanpur

geometry by comparing learning-based DOA estimation using co-prime and uniform circular microphone array.

The rest of the paper is structured as follows. Section-II introduces the signal model and co-Prime circular microphone array's architecture. This section also explains the proposed framework, which uses features extracted using co-prime and uniform array for DOA estimation and the algorithm used. Section-III compares the proposed framework with different DNN architecture and microphone array geometry, this also explains the data-set used for the learning purpose. Section-IV concludes the paper with a discussion on future work.

II. Learning based DOA Estimation using Co-Prime Circular Microphone Array

In this section, the problem of learning based DOA estimation is formulated and explained. Subsequently, a CNN based architecture is detailed for the proposed framework, which uses directional features generalised cross-coorelation with phase transfrom (GCC-PHAT) [9] using a co-prime circular microphone array.

A. Signal Model for DOA estimation

Let the signals acquired at the microphone array be $y(n,\theta) = [y_1(n,\theta), y_2(n,\theta), \ldots, y_L(n,\theta)]$, where $\theta \in [1°, 360°]$ is the angle related to the direction of arrival with respect to center of the microphone array geometry. The microphone array captured signals $y(n,\theta)$ can be mathematically represented as

$$y(n,\theta) = h(n,\theta) \circledast x(n,\theta) + v(n) \tag{1}$$

where $x(n,\theta)$ denotes the source signal due to the source located in the direction θ. $h(n,\theta) = [h_1(n,\theta), h_2(n,\theta), ..., h_L(n,\theta)]$, is the room impulse response (RIR) which depends upon the source position, each microphone position and the room dimensions and $v(n) = [v_1(n), v_2(n), ..., v_L(n)]$, is the additive background noise, which is assumed to uncorrelated for each microphone. L being the number of microphone in the array.

B. Co-Prime Geometry for Circular Microphone Arrays

Co-Prime Circular Microphone Array (CCMA) is a combination of two sparse sub-arrays, each consisting of A and B number of microphones respectively, such that A and B are co-prime. CCMA can be visualized in Figure-1, in which two sub-arrays are denoted by blue and red color respectively, with one microphone common (orange) in both, which makes the total number of microphones $L = A + B - 1$.

Assuming the reference microphone of CCMA on the x-axis the time delay between the i^{th} microphone and the centre is given by [27]

$$\tau_i = \frac{r}{c}\cos(\theta - \phi_i,) \tag{2}$$

where $i = 1, \ldots, L$, r is the radius of the CCMA, c is the speed of sound (334 m/s) and ϕ_i is given as

$$\phi_i = \begin{cases} \frac{\pi i}{A}, & i \bmod 2 = 0 \\ \frac{\pi(i-1)}{B}, & i \bmod 2 = 1. \end{cases} \tag{3}$$

.

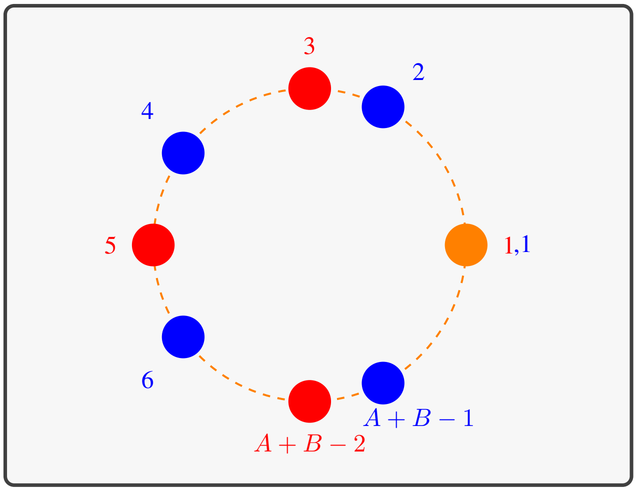

Fig. 1. Illustration of Co-prime Microphone array Geometry, which comprises of two sub-arrays depicted in red and blue, with orange as the common position point.

C. GCC-PHAT Representation for DOA Estimation

For robust DOA estimation, it is very important to estimate the TDOA between the microphones as accurately as possible. Therefore GCC-PHAT is one of the most promising approaches that can be taken into consideration for robust TDOA estimation. Considering the microphones i and j the TDOA is estimated using GCC-PHAT and given as

$$\gamma_{ij} = \frac{1}{2\pi}\int_{-\infty}^{\infty} \psi_{ij} y_i(\omega) y_j^*(\omega) e^{j\omega t} d\omega. \tag{4}$$

where γ_{ij} is the GCC-PHAT feature , $y_i(\omega)$ and $y_j^*(\omega)$ is the Fourier transform of the signal received at i^{th} and conjugate Fourier transform of the signal received at j^{th} microphone respectively. The GCC-PHAT makes use of phase transfrom as weighing factor which is defined as

$$\psi_{ij} = \frac{1}{|y_i(\omega) y_j^*(\omega)|} \tag{5}$$

The GCC-PHAT features are extracted and applied as an input for the DNN classifier, and the feature representations are shown in Figure-2 and 3 for 0 dB SNR and clean condition respectively. Each pattern corresponds to a particular angle of arrival. The same audio and the noise is used for the representation of the features for different angles. It is evident that the CCMA is less affected by the noise as the feature representation has less distortion than the UCMA features. This can be witnessed by the feature representation in the Figure-2 and 3 at lower SNR and the clean conditions, which show that the microphone array plays an important role for

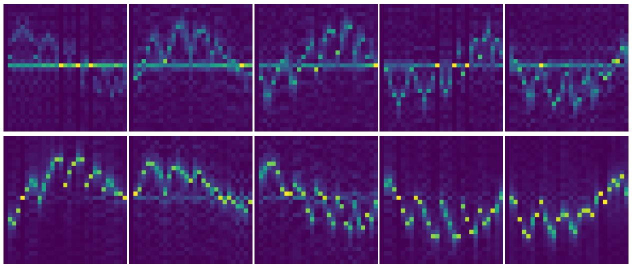

Fig. 2. Comparison of GCC-PHAT feature representation for angle from 0-240 degree with step size of 60 degree at 0 dB SNR for: (a) UCMA(above) and (b) CCMA(below)

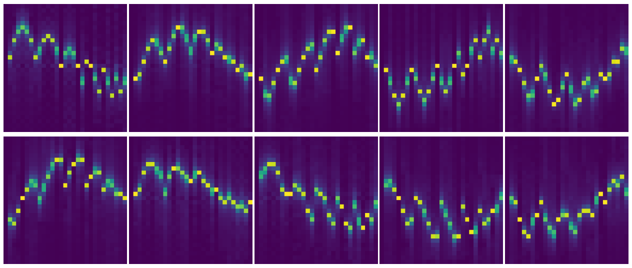

Fig. 3. Comparison of GCC-PHAT feature representation for the angle from 0-240 degree with a step size of 60 degree in clean conditions for (a) UCMA(above) and (b) CCMA(below)

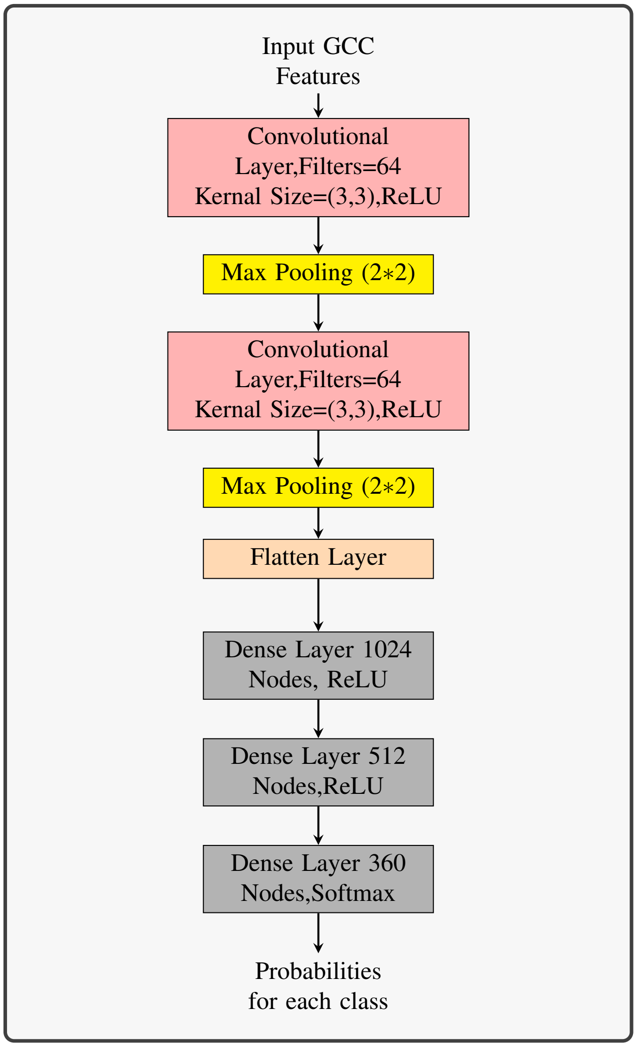
Input GCC Features
Convolutional Layer,Filters=64 Kernal Size=(3,3),ReLU
Max Pooling (2*2)
Convolutional Layer,Filters=64 Kernal Size=(3,3),ReLU
Max Pooling (2*2)
Flatten Layer
Dense Layer 1024 Nodes, ReLU
Dense Layer 512 Nodes,ReLU
Dense Layer 360 Nodes,Softmax
Probabilities for each class

Fig. 4. CNN architecture for the proposed framework for DOA classification, based on GCC-PHAT as the Input Features.

robust features extraction and so for the classification based DOA estimation.

For the DOA estimation as a classification problem, the TDOA based features, as shown in Figure-2 and 3 play a very important role. For the DNN based classification extracting discriminative features for each class plays an important task towards the accuracy. If the features correlate well with the class, the classification is accurate. The above GCC-PHAT features which are discriminative for each angle has been used as the input to the DNN based classifier.

D. DOA Estimation using CNN

The DOA as a classification problem needs discriminative features so that the DNN based classifier can learn it efficiently. Given the input as GCC-PHAT features, the final layer of the DNN based classifier generates the probabilities of the 360 classes, i.e., probabilities for each DOA angle. The overall DNN based CNN architecture used as a classifier, which is indeed used for learning the patterns of the input features, which is in the matrix format $(M \times N)$ treated as an image consists of two convolutional layers for its ability to recognize the geometrical patterns via convolution. Each convolutional layer with 64 filters with kernel size (3×3) is followed by the max-pooling layer, which is used for the dimensionality reduction having the size of (2×2) hence helping the problem of over fitting by reducing the parameters to be learned which leads to increased computational speed. The dimensions of the features after the second max-pooling layer has being fattened and supplied to the two fully connected dense layer with 1024 and 512 nodes. The final dense layer, which produces the probabilities for 360 classes, has 360 nodes with the Softmax activation function. The input GCC features goes filtering and activation step operated through convolutional kernals W as

$$q = \alpha(W * \gamma_{ij} + b) \tag{6}$$

where W is the kernal, b is the bias, γ_{ij} is the input features which is interpreted as image of dimension $M \times N$, $\alpha(\cdot)$ represents the ReLu activation function for a specific layer. The kernal is computed through an ADAM optimizer which minimizes the Sparse Categorical Cross Entropy loss between the predicted output and the target values i.e., the actual label. The output of the second CNN-layer is then fed to the dense layers. The final dense layer with 360 nodes with SOFTMAX activation function produces probabilites for each class which is given as

$$p(\theta_k = l) = \frac{e^{(q_k(l))}}{\sum_{h=1}^{H} e^{(q_k(h))}};\ l \in [0, H-1] \qquad (7)$$

Where $H = 360$ classes, $k \in [1...K]$ where K is the total number of frames from the previous dense layer. The CNN architecture is illustrated in Figure-4, which is used as the learning architecture in this work.

E. Algorithm for Learning Based DOA Estimation

Algorithm 1: Simulation for Training Dataset and Learning Methodology

```
Input: Source Signal $x_j$
Output: Microphone Signal $y_{i,j}$
Dataset Simulation for Training Phase
Room Size = 7m x 5m x 3m (in meters)
Source Distance = 4m (in meters)
$\theta \in [1°, 2°, ....., 360°]$
Microphones $\in [1, 2, ..., 8]$
for Each $j$ in $\theta$ do
    for Each $i$ in Microphones do
        RT$_{60}$ = rand (0.1, 1) (in ms)
        SNR = rand (0, 20) (in dB)
        $h_{i,j}$ = rir_generator (Room Size, Source Distance, RT$_{60}$)
        $y_{i,j}$ = h$_{i,j} \circledast x_j + a_{i,j}$
    end
    $F_j$ = GCC-PHAT (($y_j$))
end

Input: Input Features : $F = [F_1, ....., F_{360}]$
Input: Output Labels ($\theta$) = $[1°, ....., 360°]$
Output: Microphone Signal $y_{i,j}$
Training Phase
Epochs $\in [1, 2, ..., K]$
for Each $k$ in Epochs do
    $\hat{\theta}$ = CNN($F$,W$_k$)
    L = Loss ($\hat{\theta}$, $\theta$)
    W$_k$ = BackProp(L)
end

Input: Input Features : $F = [F_1, ....., F_{360}]$
Input: Trained CNN Weights: W
Output: Predicted DOA
DOA Estimation Phase
$\hat{\theta}$ = CNN(F,W)
RMSE = calculate_rmse($\hat{\theta}$, $\theta$)
MAE = calculate_mae($\hat{\theta}$, $\theta$)
```

Algorithm-1 shows the data simulation,training and DOA estimation phase.

TABLE I
ACOUSTIC PARAMETERS FOR SIMULATING TRAINING DATA

Simulated Training Dataset	
Input Speech Audio	75,600 audios(42 for each angle) of variable lenght
Room Size(m)	7m x 5m x 3m
Source Distance(m)	4m
Reverberation time (T60) (s)	0.1s to 1s (randomly chosen)
SNR(dB)	Uniformly sampled between 0dB to 20dB

TABLE II
ACOUSTIC PARAMETERS FOR SIMULATING DEVELOPMENT DATA

Simulated Development Dataset	
Input Speech Audio	5,400 Different audios(5 for each angle) of variable lenght
Room Size(m)	6m x 8m x 4m
Source Distance(m)	5m
Reverberation time (T60) (s)	0.1s to 1s (randomly chosen)
SNR(dB)	0dB, 10dB, 20dB

III. PERFORMANCE EVALUATION

This section explains the conditions taken into consideration for the generation of the dataset used to conduct the experiments. It also shows the performance comparison of the proposed framework with other existing state-of-the-art frameworks.

A. Experimental Dataset

As the learning-based algorithm requires a good amount of dataset to learn. The datasets are simulated under various room conditions, SNR levels, and RT$_{60}$ time. The data is simulated using 8-channel circular array with diameter of 20cm comprising of both the array geometry arrangement i.e. UCMA and CCMA respectively. The RIR is simulated using the Image source method (ISM) method [28], [29], [30] and convolved at various RT$_{60}$ parameter, with the clean speech taken from Librispeech database [31]. The convolved data is further augmented using additive noise taken from MUSAN noise database [32] at different SNR.

B. Experimental Conditions and Performance Measure

In Table-I and II the experimental conditions for the datasets are shown, in which the room size, RT$_{60}$ and the SNR level used for simulating the dataset is mentioned. The learning phase of the experiments requires both the training and validation datasets. Therefore the simulated data was divided into two parts. The simulated data has total of 75,600 files, from which randomly chosen 200 audios for each DOA angle which are added to 72,000 audio files, and are used for training purpose. Similarly randomly chosen 10 audios for each DOA angle which are added to 3,600 audios for validation purpose. For the development purpose, 5 audios for each DOA are simulated as per the parameters in Table-II, which add to 5,400 audios with 1,800 audios for each SNR. The performance of the proposed framework on this testing data was measured using two parameters root mean squared error(RMSE) and mean absolute error(MAE) which were calculated using (8) and (9) respectively.

TABLE III
COMPARISON OF RMSE AND MAE SCORES FOR VARIOUS ESTIMATION ALGORITHMS AND COMBINATION OF ARRAY GEOMETRY AT DIFFERENT SNR LEVELS.

Microphone Array	*Methods*	*SNR 0dB*	*SNR 10dB*	*SNR 20dB*
Root Mean Squared Error(RMSE)				
Uniform	**GCC-LS**	18.26	6.52	4.19
	DNN-SL	0.4	0.36	0.25
	CNN	0.16	0.1	0.089
Co-Prime	**GCC-LS**	12.11	4.33	3.62
	DNN-SL	0.29	0.17	0.098
	CNN	**0.14**	**0.050**	**0.036**
Mean Absolute Error(MAE)				
Uniform	**GCC-LS**	8.72	0.96	0.87
	DNN-SL	0.21	0.19	0.1
	CNN	0.13	0.093	0.081
Co-Prime	**GCC-LS**	3.57	0.84	0.73
	DNN-SL	0.16	0.11	0.093
	CNN	**0.097**	**0.032**	**0.021**

$$RMSE = \frac{1}{V}\sum_{v=0}^{V}\left[\left(\widehat{\theta}_v - \theta_v\right)^2\right] \tag{8}$$

$$MAE = \frac{1}{V}\sum_{v=0}^{V}\left(\left|\widehat{\theta}_v - \theta_v\right|\right) \tag{9}$$

where $V = 360$ whereas θ_v and $\widehat{\theta}_v$ are actual and predicted probabilites of the DOA.

C. Experimental Results

The experiments are conducted at various SNR with different combinations of microphone array geometry and neural network architecture. From Table-III, it can be clearly observed that the proposed framework, which uses CNN and Co-Prime array, performs significantly better than the other framework listed at all SNR levels using RMSE and MAE as the performance metric. Also, CNN based classification achieves a better result than the state-of-the-art LS method (GCC-LS) and the single-layer neural network (DNN-SL) [17] because of the reason that our input GCC-PHAT features are in the matrix form (M×N). Figure-5 shows the RMSE values for different array geometries for various SNR levels at each epoch of the validation set for CNN. It can also be inferred from Figure-5 that in learning-based DOA estimation, array geometry does matter. This is due to the fact of feature representations, when extracted from Co-Prime array, are more robust to noise, which can be seen in both Table-III and Figure-5. At lower SNR levels, it can be observed that when array geometry is Co-Prime, lower values of RMSE and MAE are achieved when compared with other frameworks.

To show the robustness of the CCMA when compared to UCMA for GCC-LS and CNN methods, a new set of data is generated with 15 audios for 10 DOA from 0 to 324 with a step size of 10 degrees. The source distance considered is 4.5m with SNR of 0 dB. The average error in DOA estimation is calculated using

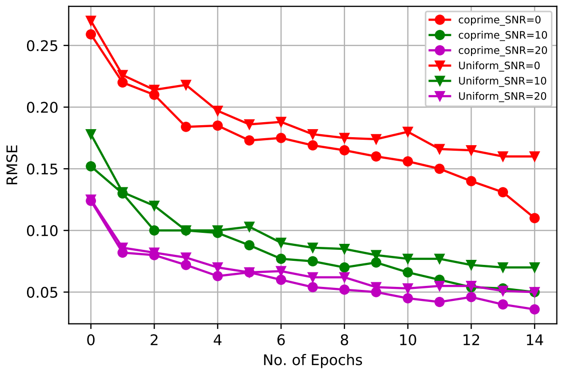

Fig. 5. Comparison of RMSE scores of Uniform and Co-Prime Circular Array for CNN method on validation dataset.

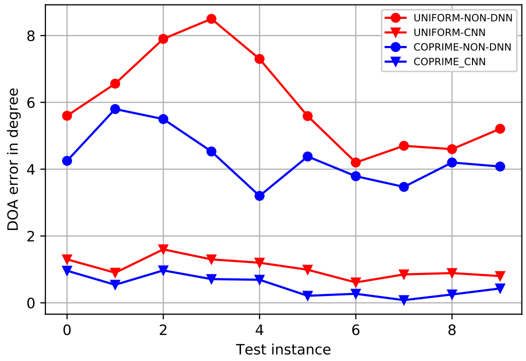

Fig. 6. Average error in DOA estimation using CCMA and UCMA for CNN and NON-DNN method.

$$E_{avg} = \frac{1}{E}\sum_{e=1}^{E}(\hat{\theta}_e - \theta_e), \tag{10}$$

where E_{avg} is the average error, and E is the total number of audios for a particular instance. Figure-6 shows the average error in DOA estimation with GCC-LS and CNN. It can be observed that the CCMA is performs better than the UCMA irrespective of the methods. It also shows the consistency in the error for the CNN based DOA estimation.

IV. CONCLUSION AND FUTURE WORK

In this work, the significance of using a co-prime arrangement over the uniform arrangement of the array with learning-based DOA estimation is discussed. The paper also shows the superiority of the proposed framework with traditional signal processing algorithms such as Least Squares (LS). The robust representation of features at lower SNR using Co-Prime arrangement is leveraged, and it is observed that for learning based DOA estimation, array geometry is indeed important because feature representation is dependent on it.

The future work can be extended to generalizing the model on various types of acoustic environments, like with different rooms and positions of source and microphone. This arrangement can be further extended to the Multi-Channel speech enhancement task, which needs to be explored.

References

[1] 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 *2012 7th IEEE Conference on Industrial Electronics and Applications (ICIEA)*. IEEE, 2012, pp. 1853–1857.

[2] S. Zhao, E. S. Chng, N. T. Hieu, and H. Li, "A robust real-time sound source localization system for olivia robot," in *2010 APSIPA annual summit and conference*, 2010.

[3] M. Wölfel and J. McDonough, *Distant speech recognition*. John Wiley & Sons, 2009.

[4] 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 *Proc. REVERB challenge workshop*, 2014, p. o2.

[5] R. O. Schmidt, "A signal subspace approach to multiple emitter location and spectral estimation." 1982.

[6] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE transactions on antennas and propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[7] X. Zhong and J. R. Hopgood, "A time–frequency masking based random finite set particle filtering method for multiple acoustic source detection and tracking," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 12, pp. 2356–2370, 2015.

[8] Y. A. Huang, J. Benesty, and J. Chen, "Time delay estimation and source localization," in *Springer Handbook of Speech Processing*. Springer, 2008, pp. 1043–1063.

[9] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE transactions on acoustics, speech, and signal processing*, vol. 24, no. 4, pp. 320–327, 1976.

[10] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays*. Springer, 2001, pp. 157–180.

[11] Y. Huang, J. Benesty, G. W. Elko, and R. M. Mersereati, "Real-time passive source localization: A practical linear-correction least-squares approach," *IEEE transactions on Speech and Audio Processing*, vol. 9, no. 8, pp. 943–956, 2001.

[12] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 1997, pp. 375–378.

[13] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*. Springer Science & Business Media, 2008, vol. 1.

[14] P. Stoica and K. C. Sharman, "Maximum likelihood methods for direction-of-arrival estimation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 7, pp. 1132–1143, 1990.

[15] S. Delikaris-Manias, D. Pavlidi, A. Mouchtaris, and V. Pulkki, "Doa estimation with histogram analysis of spatially constrained active intensity vectors," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 526–530.

[16] A. H. Moore, C. Evers, and P. A. Naylor, "2d direction of arrival estimation of multiple moving sources using a spherical microphone array," in *2016 24th European Signal Processing Conference (EUSIPCO)*. IEEE, 2016, pp. 1217–1221.

[17] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, 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.

[18] S. Chakrabarty and E. A. Habets, "Broadband doa estimation using convolutional neural networks trained with noise signals," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2017, pp. 136–140.

[19] C. Xu, X. Xiao, S. Sun, W. Rao, E. S. Chng, and H. Li, "Weighted spatial covariance matrix estimation for music based tdoa estimation of speech source." in *INTERSPEECH*, 2017, pp. 1894–1898.

[20] W.-K. Ma, T.-H. Hsieh, and C.-Y. Chi, "Doa estimation of quasi-stationary signals with less sensors than sources and unknown spatial noise covariance: A khatri–rao subspace approach," *IEEE Transactions on Signal Processing*, vol. 58, no. 4, pp. 2168–2180, 2009.

[21] M. Agrawal and S. Prasad, "A modified likelihood function approach to doa estimation in the presence of unknown spatially correlated gaussian noise using a uniform linear array," *IEEE transactions on signal processing*, vol. 48, no. 10, pp. 2743–2749, 2000.

[22] J. Xie, Z. He, H. Li, and J. Li, "2d doa estimation with sparse uniform circular arrays in the presence of mutual coupling," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, no. 1, p. 127, 2011.

[23] J. Zhao and C. Ritz, "Investigating co-prime microphone arrays for speech direction of arrival estimation," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2018, pp. 1658–1664.

[24] ——, "Co-prime circular microphone arrays and their application to direction of arrival estimation of speech sources," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 800–804.

[25] F. Dong, Y. Jiang, Y. Yan, Q. Yang, L. Xu, and X. Xie, "Direction-of-arrival tracking using a co-prime microphone array: A particle filter perspective," *Applied Acoustics*, vol. 170, p. 107499, 2020.

[26] J. Zhao and C. Ritz, "Semi-coprime microphone arrays for estimating direction of arrival of speech sources," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2019, pp. 308–313.

[27] J. Benesty, J. Chen, and I. Cohen, *Design of Circular Differential Microphone Arrays*. Springer, 2015, vol. 12.

[28] 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, 1979.

[29] E. A. Lehmann and A. M. Johansson, "Diffuse reverberation model for efficient image-source simulation of room impulse responses," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1429–1439, 2009.

[30] ——, "Prediction of energy decay in room impulse responses simulated with an image-source model," *The Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 269–277, 2008.

[31] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[32] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," 2015, arXiv:1510.08484v1.