Coprime Microphone Arrays for Estimating Speech Direction of Arrival Using Deep Learning

Jiahong Zhao and Christian Ritz
School of Electrical, Computer and Telecommunications Engineering, University of Wollongong, NSW, Australia
E-mail: jz262@uowmail.edu.au and critz@uow.edu.au

Abstract— **This paper investigates deep neural network (DNNs) applied to coprime microphone arrays (CPMAs) and semi-coprime microphone arrays (SCPMAs) for direction of arrival (DOA) estimation of speech signals. Existing research has shown that the coprime arrangement increases the operating frequency of conventional uniform linear arrays (ULAs) by interleaving two uniform sub-arrays with different spacing. The SCPMA extends this arrangement and further increases the operating frequency, above which interfering signals are largely amplified in the recording and lead to confusion with the desired source. As a result, both types of coprime geometries improve the beampattern, array gain and DOA estimation results compared to the ULA. However, large side lobes may still occur in the beampattern of the two coprime arrangements, resulting in degraded DOA estimates using conventional beamforming-based approaches in an adverse environment. The proposed approach alternatively utilises deep learning (DL) to estimate speech DOAs using coprime microphone arrays. Experimental results evaluating the accuracy under different levels of noise using the mean absolute error (MAE) and root mean square error (RMSE) of the DOA estimate indicate satisfactory performance of the proposed method.**

I. INTRODUCTION

Microphone arrays have been proposed for decades, which utilise multiple microphones in specific arrangements for recording and have been employed in many applications such as direction of arrival (DOA) estimation [1], [2], acoustic source separation [3] and robust sound source tracking [4]. Among all types of acoustic signals, speech has characteristics of being time-varying and broadband, leading to difficulties to developing solutions for the aforementioned applications.

Traditional microphone arrays have a uniform arrangement, in which all microphones are equispaced and can be either uniform linear arrays (ULAs) or uniform circular arrays (UCAs) [5] in a two-dimensional (2D) space. By modifying the interelement spacing, increasing the number of sub-arrays or scattering array elements, improved geometries have been proposed and applied, such as the differential microphone arrays (DMA) [6], co-prime microphone array (CPMA) [7], [8], semi-coprime microphone array [9], [10] and ad-hoc microphone array [11].

A commonly-used method to evaluate the performance of microphone arrays is DOA estimation. Compared with its application to signals at a single frequency point, DOA estimation of broadband signals, e.g. speech, is more challenging. When analysing such signals of multiple frequency bands, a crucial issue that is usually encountered is spatial aliasing, occurring above the operating frequency of conventional ULAs that is based on the inter-element spacing. When there is spatial aliasing, a few grating lobes will be present in the beampattern of microphone arrays, having the same amplification of the main lobe and thus significantly interfering the reception of desired signals. The coprime arrangement has been shown to possess the capability of cancelling spatial aliasing in [7], [8], [12] with two co-located sub-arrays without changing the aperture and number of microphones. Although the grating lobes were successfully removed within the frequency range of typical speech signals, large side lobes may still exist in the coprime beampattern. The SCPMA mitigated these large side lobes and obtained satisfactory array gains using both the product processor and min processor [9], [10]. However, the speech DOA estimation results using steered response power - phase transform (SRP-PHAT) presented unexpected large errors under high noise and strong reverberation. Thus, alternative DOA estimation approaches are needed to further explore the advantage of coprime microphone arrays.

In [10], the SRP-PHAT algorithm calculated the summation of the generalised cross-correlation - phase transform (GCC-PHAT) of all microphone pairs of the array, which has been demonstrated to be robust to noise and reverberation [13]. Seeing that the SRP-PHAT did not work as expected for coprime-based microphone arrays yet, taking one step back and investigating alternative approaches to process the GCC-PHAT is a potential direction to increase the robustness to adverse environments. With the rapid development of deep learning (DL), researchers have been utilising GCC-PHAT as a feature to predict the DOA of sound sources. The work in [14] started to use GCC-PHAT for training, where speech DOAs were classified by a multilayer perceptron (MLP), showing accurate results in most interfering scenarios. After that, the network structure has evolved into more sophisticated networks, including deep neural networks (DNNs) with additional encoders and hidden layers [15], CNNs for broadband DOA estimation [16] and sound source localisation in a multipath environment [17], convolutional recurrent neural network (CRNN) for jointly localising sound events and detecting overlapping sources [18], etc. Most neural networks in this work use non-casual recurrent layers and are not suitable for real-time applications as mentioned in [4], which proposed a three-dimensional (3D) CNN training a new power map,

leading to robust sound source tracking.

This paper explores the use of GCC-PHAT to CPMA and SCPMA for robust DOA estimation in adverse environments. A 2D CNN structure is designed to start this exploration. Firstly, GCC-PHAT is extracted as a 2D feature from CPMA and SCPMA recordings, which have been preprocessed by removing all silence and unvoiced segments that do not contain useful DOA information for training. The wideband signals in IEEE corpus [19] is chosen to make comparisons with [10], which provides the ground truth of voiced segments. Different levels of noise are added to the speech-only recordings in simulation. Secondly, we model the DOA estimation of speech sources as a classification task with each class being an individual DOA of concern. The proposed CNN treats the input GCC-PHAT as one-channel gray-scale images and directly learns the underlying non-linear relationships between GCC-PHAT and the corresponding DOA. The outputs of DOA are expected to be robust to high noise in test cases under similar conditions with the training set.

The main contributions of this paper are: 1) proposing the improved GCC-PHAT formulation for SCPMA to make full use of the robustness of GCC-PHAT to interferences for DOA estimation; 2) designing a DL approach based on 2D CNN to train the feature of GCC-PHAT to investigate coprime array arrangements, named CoprimeNet, which brings high accuracy with stable error distributions for speech DOA estimation under high noise using ULA, CPMA and SCPMA; 3) evidencing advantages of the SCPMA geometry in combatting high noise when estimating DOA of speech signals. Note that although the proposed method is used for coprime microphone arrays, the application is not limited to certain array geometries, as the feature of GCC-PHAT and designed CNN structure are both universal. Moreover, as the size and number of included scenarios of the training data increases, the proposed DL approach tends to bring more reliable DOA estimates and has the potential to be applied in real-world environments.

The remainder of this paper is organised as follows. Section II reviews the coprime arrangements discussed in this paper, and Section III customises a 2D CNN structure trained with GCC-PHAT used for robust speech DOA estimation. Experiments of the proposed method are conducted in Section IV, before evaluating their results and comparing them with literatures. This paper is concluded in Section V, with future work also outlooked.

II. Mathematical Model for Coprime and Semi-Coprime Microphone Arrays

A. Models of the CPMA and SCPMA

The traditional CPMA is a type of sparse array. It interleaves two uniform linear sub-arrays, and their inter-element spacing and numbers of microphones are coprime related, where the only positive number that divides both is 1. A typical microphone arrangement of the CPMA is delineated in Fig. 1 [8]. There are three parameters needed to define a CPMA, including the numbers of microphones M and N ($M > N$), and the unit of inter-element distance d. The two sub-arrays share

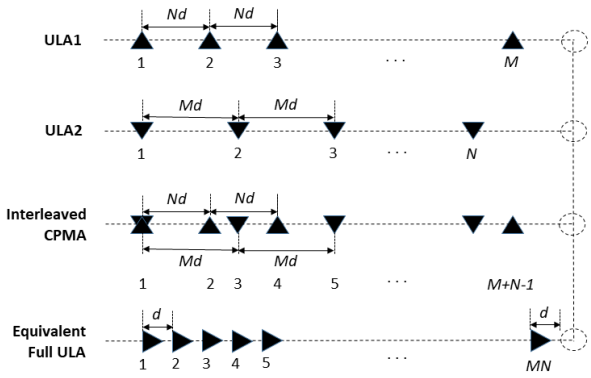


Fig. 1 The geometry of conventional CPMAs

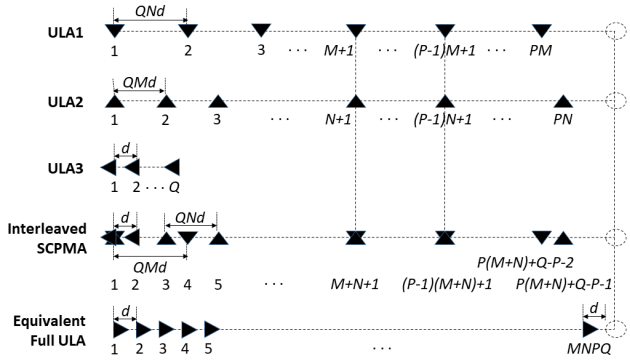


Fig. 2 The generalised structure of the SCPMA

the first microphone and have identical array apertures, leading to an overall coprime array with $M + N - 1$ microphones. To distinguish the expression of microphone numbers of CPMA from that of SCPMA, the M, N and d for CPMA are written as M_c, N_c and d_c, respectively, in the rest of the paper.

The SCPMA extends the idea of coprime array and includes an extra short sub-array to improve the overall beampattern, leading to three uniform linear sub-arrays [9]. The generalised SCPMA structure is determined by five parameters, including M, N, P, Q and d, which are illustrated in Fig. 2 [10]. Here the M and N are a pair of coprime numbers, whereas the numbers of microphones of any two SCPMA sub-arrays are not necessarily coprime related, particularly for the first two sub-arrays in Fig. 2, the coefficient P is another positive number that divides both M and N. The three SCPMA sub-arrays have PM, PN and Q microphones, separately, with their inter-element spacing defined as QNd, QMd and d, respectively. The equivalent virtual full ULA to SCPMA shows an inter-microphone distance of d, which is also the unit of that for the first two sub-arrays. Q is usually small in the SCPMA design, so the third sub-array is quite short. The three sub-arrays share the first microphone, and the apertures of them are all the same with that of the overall SCPMA and equivalent full ULA, with a virtual microphone located at the rightmost of all of them. The total number of microphones of SCPMA can be calculated as $U = P(M + N) + Q - P - 1$ [10].

To model their signal recording mathematically, K uncorrelated acoustic sources are assumed to impinge on the

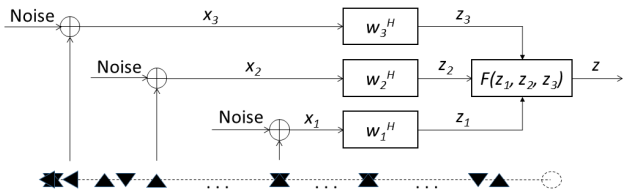


Fig. 3 Common workflow of processors for sub-array signal processing

microphone array, propagating as plain waves in a 2D space at the speed of sound (c = 343 *m/s*) from θ_i (i = 1, 2, … , K). The signal recording is then modelled as [10]

$$y_u(t) = \sum_{i=1}^{K} h_{u,i}(t) * s_i(t) + v_u(t), \quad (1)$$

where u = 1, 2, …, U. $y_u(t)$ is the time domain output of each individual microphone, which is the recorded signal. $h_{u,i}(t)$ is the room impulse response (RIR) of source i received by microphone u. $s_i(t)$ and $v_u(t)$ are the original signal of the ith source and additive noise to the uth microphone, respectively. The individually recorded signals are combined through certain algorithms for each sub-array, and the overall output combines all sub-array outputs using a processor.

B. Processors and Operating Frequencies

The common workflow is illustrated in Fig. 3, which uses SCPMA as an example. The processor F is a function to combine weighted sub-array outputs z_i into the overall microphone array output z, which is expressed as

$$z = F(z_1, z_2, \dots, z_A), \quad (2)$$

where A is the number of sub-arrays, i.e. A = 2 for CPMAs and A = 3 for SCPMAs. Each weighted sub-array output z_i is obtained through

$$z_a = w_a^H x_a \quad (3)$$

where H denotes the conjugate-transpose operation, w_a (a = 1, 2, …, A) represents the beamforming weight for the ath sub-array, and x_a is the sub-array output before the weighting calculation. The processor discussed in this paper is the product processor, which multiplies weighted sub-array outputs and matches the feature for training to be formulated in Section III.

The proposal of the coprime-based array geometries mainly aims at addressing the spatial aliasing problem, occurring above the operating frequency of microphone arrays. According to the spatial Nyquist sampling theorem, when the inter-microphone spacing δ is larger than half of the wavelength λ of sound sources, i.e. $\delta > \lambda / 2$, there will be ambiguity in distinguishing the desired source from signals coming from other directions. The threshold frequency that satisfies $\delta = \lambda / 2$ is defined as the operating frequency. For a ULA with N_0 microphones, it is formulated as

$$f_{op\_ULA} = \frac{c}{2\delta} = \frac{cN_0}{2L_{ULA}}, \quad (4)$$

where L_{ULA} is the aperture of ULA. The operating frequency of CPMA is identical with that of its equivalent full ULA, shown as [7]

$$f_{op\_CPMA} = \frac{cM_cN_c}{2L_{CPMA}}, \quad (5)$$

where L_{CPMA} is the CPMA's aperture. Similarly, the aperture of SCPMA is derived as

$$f_{op\_SCPMA} = \frac{cMNPQ}{2L_{SCPMA}}, \quad (6)$$

where L_{SCPMA} is the aperture of SCPMA.

III. Convolutional Neural Network Trained with GCC-PHAT

A. GCC-PHAT as Features for Training

It has been widely recognised that GCC-PHAT contains important DOA information and can lead to robust DOA estimation under certain levels of noise and reverberation [13]. In our previous work [10], SRP-PHAT was utilised to add GCC-PHATs of all combinations of microphone pairs together at each steering angle before finding the maximum summation to define the estimated DOA. This approach of using GCC-PHAT did not lead to highly accurate DOA estimates as expected under high noise and strong reverberation. This paper further explores the potential of GCC-PHAT for robust speech DOA estimation using coprime-based array geometries through DL, which will be discussed in Section III-B.

Traditionally, GCC-PHAT is used for ULAs, and its calculation starts from the cross-correlation for one microphone pair in the frequency domain, which can be written as

$$\vartheta_{y_1y_2}(f) = E[Y_1(f)Y_2^*(f)] \quad (7)$$

where * denotes the complex conjugation, and $E[\cdot]$ calculates the mathematical expectation. $Y_i(f)$ $(i = 1, 2)$ is the frequency-domain output of an individual microphone in the selected pairs. The GCC transforms $\varphi_{y_1y_2}(f)$ to time domain by using the inverse discrete-time Fourier transform (IDTFT) after applying a weighting function. If choosing the PHAT weighting that cancels the amplitude of $\varphi_{y_1y_2}(f)$ and only convey the phase information, the resulting formulation will calculate the GCC-PHAT, which is

$$\varphi_{y_1y_2}(\tau) = \int_{-\infty}^{+\infty} \frac{\vartheta_{y_1y_2}(f)}{|\vartheta_{y_1y_2}(f)|} e^{j2\pi f\tau} df \quad (8)$$

Based on (8) for one microphone pair, all GCC-PHAT values for a conventional ULA can be found by selecting all possible microphone pairs. If it is an N_0-element ULA, the number of microphone pairs will be $\frac{N_0(N_0-1)}{2}$, which does not count repeated microphone pairs and is also the number of GCC-PHAT for the ULA.

For CPMA, the GCC-PHAT has been formulated in [20]. The method outperforms the formulation of GCC-PHAT in [10], which calculates GCC-PHAT for each uniform sub-array and does not utilise the coprime geometry. The GCC-PHAT used for CPMA has the same equation with (8) for one microphone pair, whereas the two microphones are restricted to be from different sub-arrays. Considering a CPMA having

M_c and N_c microphones in each sub-array, separately, the overall numbers of microphone pairs will be $M_cN_c - 1$, and the only removed "pair" is formed by the first microphone of both sub-arrays, which is actually the same one. This approach exactly matches the definition of GCC-PHAT to the CPMA using the product processor.

In this paper, we propose a similar definition of GCC-PHAT for SCPMA to fully employ the robustness of GCC-PHAT to adverse environments for speech DOA estimation. The corresponding GCC-PHAT calculation in [10] suffers from the same problem with CPMA and does not utilise the semi-coprime arrangement. Here we remain the GCC-PHAT equation for one microphone pair the same as (8), whereas the selection of all pairs is much more complex. The selection in CPMA is straightforward, with all microphones from one sub-array pairing with all microphones from the other before removing the pair of the shared first microphone. The difficulty for SCPMA lies in choosing two microphones from three sub-arrays. We define the process to select microphones from any two of the three sub-arrays except for pairs of shared microphones and repeated pairs. Firstly, considering the first two sub-arrays in Fig. 2, the number of all microphone pairs without removing extra ones is $PM{\cdot}PN$. There are P repeated microphones between ULA1 and ULA2, so the number of repeated pairs equals the total number of microphones of a P-element ULA, which is $\frac{P(P+1)}{2}$. Note that the self-pairing cases are also removed in this calculation. Thus, the overall pair number for the first sub-arrays is $P(PMN - \frac{P+1}{2})$. Similarly, we obtain the pair number for ULA1 and ULA3 as $PM(Q - 1)$, and also that for ULA2 and ULA3 as $P(N - 1)(Q - 1)$. In summary, we give the total number of unrepeated microphone pairs of SCPMA as follows.

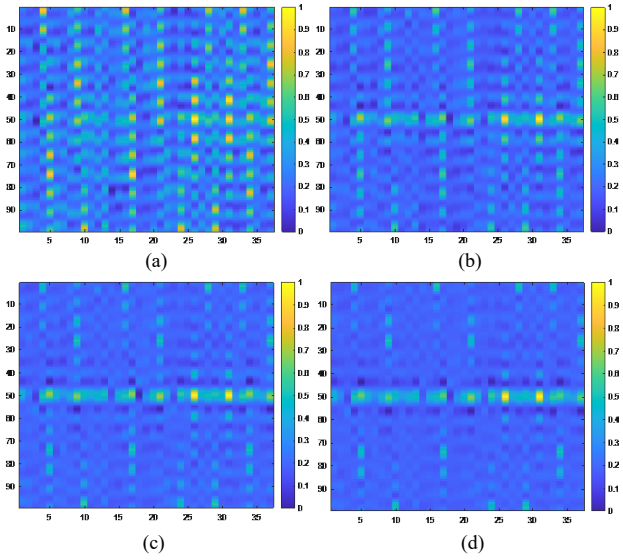


Fig. 4 Examples of the proposed GCC-PHAT calculation for SCPMA. Room dimensions: 10 × 12 × 5 m^3; source-array distance: 7m; sampling frequency: 25 kHz; DOA: 55 degree; array aperture: 0.8 m; normalised to [0, 1]; SNR: 10 dB (a), 20 dB (b), 30 dB (c), 40 dB (d).

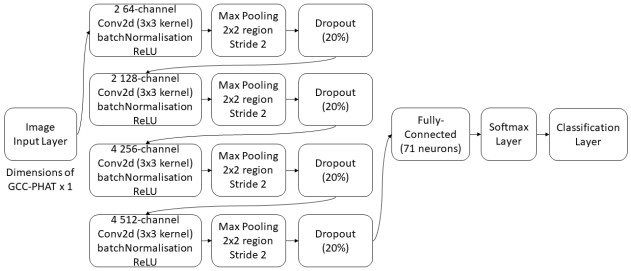


Fig. 5 The proposed 2D CNN structure for speech DOA estimation

$$D_{SCPMA} = P(PMN - \frac{P+1}{2} + M(Q-1) + (N-1)(Q-1) \quad (9)$$

For example, letting $M = 3$, $N = 2$, $P = 2$ and $Q = 3$, which forms a 10-element SCPMA. The pair number D_{SCPMA} will be 37. This is also the setting we will use in Section IV, and its example GCC-PHAT plots are shown in Fig. 4.

It can be observed that the four GCC-PHAT figures for the 10-element SCPMA shows similar patterns, and the difference in colour brightness is caused by different levels of noise. The 10 dB plot in Fig. 4 (a) shows the best contrast among all of the four, which will be easier to distinguish the underlying pattern and indicates potential robustness of GCC-PHAT to high noise.

B. CNN-based Robust DOA Estimation

To directly map the feature of GCC-PHAT to DOA estimates, a CNN structure is designed in this section to model the DOA estimation as a classification task, with each class corresponding to one concerned DOA. Inspired by VGG-19 [21], which has been widely used for large-scale image recognition, the proposed CNN structure explores the underlying information in GCC-PHAT patterns step by step using a series of convolutional layers (Conv2d) with 3-by-3 kernels, which is illustrated in Fig. 5. There are four blocks composed of a few convolutional layers, a max pooling layer and a dropout layer. The size of the overall data is reduced by 4 times after each block, which is then increased by twice through doubling the channel number of convolutional layers in the next block. For example, the GCC-PHAT is firstly input to two 64-channel Conv2d, before being processed by a max pooling layer scanning 2-by-2 regions with a stride of 2. The max pooling operation leads to data with half of both the height and width of the input image, so the output size of the first dropout becomes 1/4 of the original GCC-PHAT. The next layer uses two 128-channel Conv2d, which double the channel number of the previous Conv2d, so the size of data increases by twice. This type of multi-channel CNN structure with pooling layers have been shown to well represent high-level features of the input [4]. This paper utilises the structure to gradually looks into GCC-PHAT from general DOA information to deeper one, which is beneficial to make full use of features contained in GCC-PHAT.

In addition, each of the Conv2d is followed by a batch normalisation layer to accelerate the training process and avoid the vanishing gradient problem [22], and the activation function of all Conv2d in this paper is chosen as ReLU to

TABLE I
EXPERIMENTAL MICROPHONE ARRAY COEFFICIENTS

Type of array	Number of microphones	Aperture (m)	f_{op} (Hz)
SCPMA	10	0.8	7717.5
CPMA	10	0.8	6431.3
ULA	10	0.8	2143.8

TABLE II
SIMULATION SETTINGS

Sampling frequency (F_s)	25 kHz
Frequency bin number for FFT	625
Frame duration	200 ms
Frame overlap	50%
Azimuthal range	55˚ - 125˚
Azimuthal resolution	1˚ (71 classes)
Room dimensions	$10 \times 12 \times 5$ m^3
Noise levels (SNRs)	{10, 20, 30, 40, ∞} dB
Ground truth DOAs (S_1, S_2, S_3)	{107˚, 76˚, 56˚}
Source-array distance	7 m
Speed of sound (c)	343 m/s

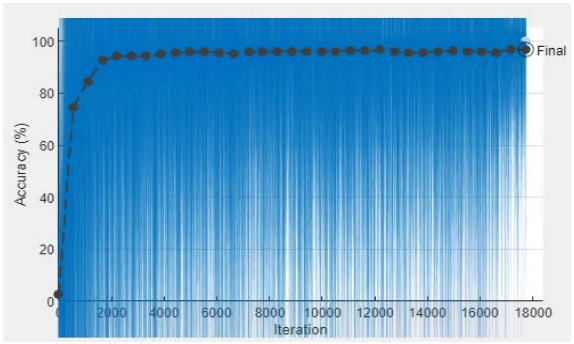
possess the capability of modelling nonlinearity, which also helps with eliminating the vanishing gradient issue and mitigates overfitting [23]. The dropout layer at the end of each block randomly selects 20% of all neurons to neglect in each iteration. The strong ability of avoiding overfitting brought by the dropout layer was evidenced by a well-known network for image classification, the AlexNet [24]. In theory, the deeper the network is, the better the input feature will be represented, whereas we find that the number of designed blocks of convolutional layers is limited. The reason lies in that after a few blocks, the dimension of the data can be reduced to 1 and cannot be further convoluted. Lastly, a fully-connected layer, a softmax layer and a classification layer is designed to output final results, which is a commonly-used approach. Extra fully-connected or dense layers that usually follow the CNN layers in literatures are not used, as multiple layers of CNN already have excellent capabilities of feature representation, and the extra computational load involved by the fully-connected layer is not desired in this design. By training a prepared dataset properly, the network output in the test phase will directly be the estimated DOA. We name the proposed structure for speech DOA estimation using coprime arrangements as CoprimeNet.

IV. RESULTS AND DISCUSSION

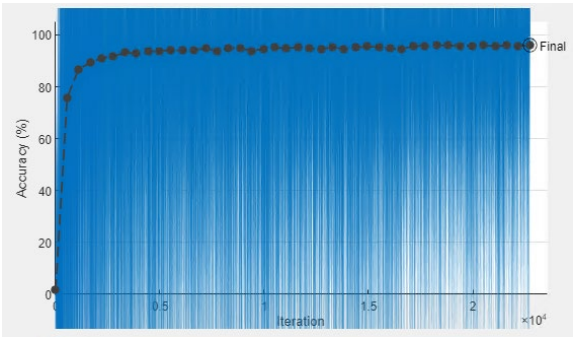
A. Experimental Configurations

Experiments are conducted to evaluate the speech DOA estimation performance of the proposed approach for the CPMA and SCPMA. The configurations of these as well as the comparative ULA are shown in Table I, which is exactly the same with [10] in order to make comparisons with the results in that work. All microphone arrays have 10 elements, which means the four parameters of SCPMA are set as $P = 2$, $Q = 3$, $M = 2$, $N = 3$, and the two for CPMA are $M_c = 5$, $N_c = 6$, respectively. Looking at the operating frequencies of the three microphone arrays of an 0.8 meter aperture, the ULA is expected to suffer from spatial aliasing within the frequency range of the chosen wideband recordings, and CPMA and SCPMA will not have this problem when estimating the DOA of speech sources [10].

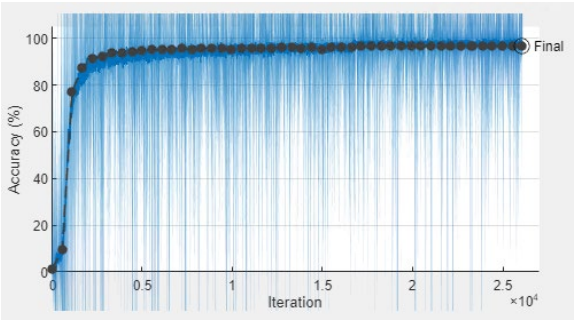
The settings of preparing the dataset used for CoprimeNet are listed in Table II. The raw speech recordings are all 720 sentences from the IEEE corpus (wideband), and they are sampled at 25 kilohertz [19]. As the silence and unvoiced segments within a recording will not contain useful DOA information for training, they are first removed using the ground truth provided by the corpus. Subsequently, the speech-only signals are convolved with RIRs simulated by using the image method [25]. For the initial evaluation in this paper, the room dimensions are 10 m × 12 m × 5 m (width × length ×



(a)



(b)



(c)

Fig. 6 Training progresses using ULA (a), CPMA (b) and SCPMA (c); blue (light colour) curve: training accuracy, black dots: validation accuracy

height), and the left-most microphone is placed at the position of 4 m along the width dimension. All microphones are located at 2 m long and 2 m high. During simulation, we find that the RIR for certain DOAs fail to be generated, leading to potential imbalance of the training data. To avoid that, we choose a concerned DOA range from 55 degree to 125 degree for balanced data generation. In addition, the source-array distance is set to a fixed value of 7 m, with same ground truth DOAs with [10] at a resolution of 1 degree. After the recording process in a simulated room, we add different levels of white noise to the signals, and the signal-to-noise ratio (SNR) ranges from 0 dB to 40 dB. The symbol of infinity in Table II represents clean source with no additive noise. We then transform the recordings to the short-term frequency domain by utilising fast Fourier transform (FFT) with 625 frequency bins, which is performed on 200 ms Hamming windowed frames with an 50% overlap. For each resulting frame, one GCC-PHAT is calculated, before averaging all GCC-PHATs for the same recording to obtain an overall feature, which is a representative of the recording to be input to CoprimeNet. The accuracy of speech DOA estimation is measured through two metrics, including the mean absolute error (MAE) and root mean square error (RMSE).

In the training phase, we use Matlab 2021a to implement the network structure. The mini-batch size is set to 128, and the maximum epoch is 80. As the training process can be sensitive to the parameter of learning rate, we use a piecewise schedule to configure it to avoid overfitting. The initial learning rate is 0.01, which will be multiplied by 0.1 every 30 epochs. All data are shuffled at the beginning of every epoch, and the validation frequency is also set to one epoch. The validation patience is 10 epochs, which means the network training automatically stops if the validation loss was larger than or equal to the previously smallest loss for 10 epochs. We use 675 of the 720 wideband sentences in [19] for training and validation, among which 80% of the generated recording is used for training. In this paper, we add noise at random levels selected from the five possibilities as shown in Table II to each recording, and then we repeat this process until there are totally 1251 GCC-PHATs generated, 1000 for training and 251 as the validation set. In this sense, the data incorporating each investigated level of noise are included and their numbers are close, so overfitting will be avoided in theory. The training progress of the three microphone arrays in this paper are shown in Fig. 6. The three training progresses all demonstrate accurate validation accuracy at the "Final" point, which are 96.89% for ULA, 95.88% for CPMA and 96.81% for SCPMA, respectively. The high accuracy indicates that GCC-PHAT is a favourable feature generally for all types of microphone arrays in this paper. The total number of iterations needed for training in each sub-plot is different, which is influenced by different complexity of GCC-PHAT for each microphone array.

In addition, it can be observed in Fig. 6 that the training accuracy plotted as blue (light colour) curves shows lots of spikes for all three microphone arrays, whereas through Fig. 6 (c), we can also find there is a smoothed blue curve that is almost overlapped with the dashed line connecting black dots.

TABLE III
SPEECH DOA ESTIMATION RESULTS USING DEEP LEARNING

SNR	10 dB	20 dB	30 dB	40 dB	50 dB
	MAE				
ULA	0.1332	0.0141	0.0012	0.0003	**0**
CPMA	0.1436	0.0318	0.0043	0.0012	0.0003
SCPMA	**0.1209**	0.0165	0.0034	0.0006	**0**
	RMSE				
ULA	0.3854	0.1237	0.0350	0.0175	**0**
CPMA	0.3986	0.1784	0.0655	0.0350	0.0175
SCPMA	**0.3582**	0.1286	0.0580	0.0247	**0**

The same smoothed blue curves occur as well in Fig. 6 (a) and Fig. 6 (b), but due to the large number of spikes in them, they cannot be easily found. The unsmoothed training accuracy for SCPMA in Fig. 6 (c) presents much less spikes and fluctuations than the other two, which indicates a stable training progress and that sufficient information of the DOA is well represented in the input feature. Note that the spikes must appear in the plot, which might be due to detailed captures of the training process in Matlab and do not affect the success of training. As the training data is synthesised with different levels of noise, the GCC-PHAT feature for SCPMA formulated in this paper shows great potential for DOA estimation robust to noise.

In the testing stage, the remaining 46 sentences of the total 720 are simulated with additive noise at all possible levels of SNR, leading to five folders of GCC-PHATs with each having 46 stored test data. Thus, we can evaluate DOA estimation results for each SNR level separately. Note that in order to make direct comparisons with the work in [10], three fixed DOAs are chosen to simulate the speakers in this paper, which are exactly the same with [10]. Future work will include a thorough comparison between the two approaches considering speech sources from random angles in a specified range.

B. Performance Evaluation of Speech DOA Estimation

After estimating DOAs of speech sources using the trained network and all test dataset, the results are listed in Table III. All DOA estimates in the table are shown to be accurate, which supports the argument that GCC-PHAT contains sufficient DOA information that is robust to noise and is a suitable feature for the usage in deep learning. The most accurate DOA estimates are from ULA and SCPMA under noise at 50 dB SNR, showing an 0 for both the MAE and RMSE. As for scenarios from 20 dB to 40 dB, the 10-microphone ULA presents closest results to the DOA ground truth, followed by the SCPMA with slight differences. When the SNR level is 10 dB, the most noisy case in this paper, the SCPMA leads to the highest accuracy, higher than that of ULA by around 9% and 7% in terms of MAE and RMSE, separately. This result evidences the advantage of the semi-coprime array geometry in resisting high noise, and the proposed GCC-PHAT formulation for SCPMA successfully explore such advantages for robust speech DOA estimation. The spatial aliasing expected to occur for ULA does not affect the accuracy of its DOA estimates, owing to the strength of deep learning in extracting underlying patterns of the input. The CPMA shows the highest error in all

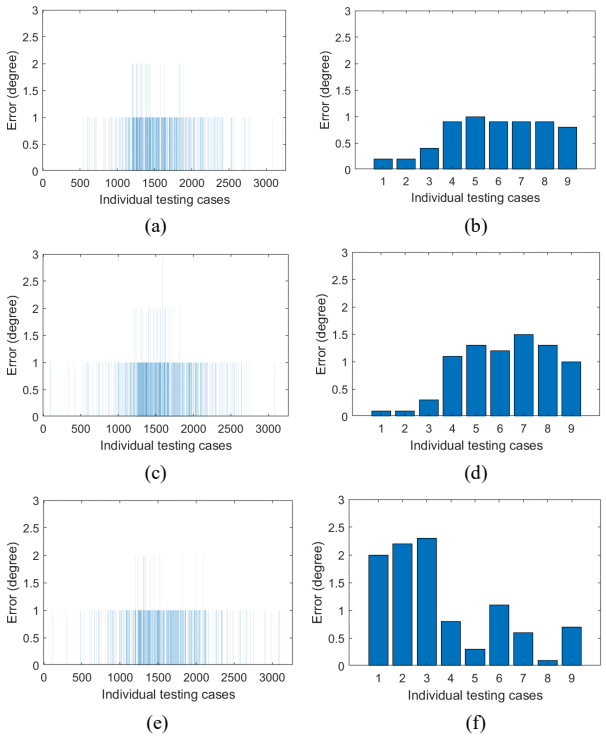


Fig. 7 Distributions of errors for 10 dB SNR applying: CoprimeNet to ULA (a), SRP-PHAT to ULA (b), CoprimeNet to CPMA (c), SRP-PHAT to CPMA (d), CoprimeNet to SCPMA (e), SRP-PHAT to SCPMA (f).

cases, which might be due to its limited number of pairs and thus small-sized GCC-PHAT. This is also the reason why we have to reduce the five convolutional blocks in [21] to four for all microphone arrays in this paper, as mentioned in Section III-B. Additionally, the slightly lower final validation accuracy of CPMA than the other two suggests less precise DOA estimates, which is also due to the limited and irregular dimensions of GCC-PHAT. We will look at alternative network structures or array configurations for solving this problem in the future.

Furthermore, we compare the experimental results in this paper with those in [10] to show the improvement in utilising SCPMA for speech DOA estimation. As we use the ground truth of voiced segments to truncate recordings which is not done in [10], it is not reasonable to directly compare the accuracy results. Instead we would plot the distribution of errors under high noise in these two papers to make comparisons, which can be found in Fig. 7. The left column of sub-plots Fig. 7 (a), (c), (e) are error distributions using the proposed CoprimeNet, and the right column Fig. 7 (b), (d), (f) are distributions of errors in [10] using SRP-PHAT. All results are simulated at 10 dB SNR. It can be found that by using the deep learning method, the errors using all three microphone arrays are concentrated, and the majority of errors have the same value. Although the number of errors in [10] is less, and the superior error distribution in this paper is partially due to the resolution of CoprimeNet is 1 degree which is lower than the 0.1 degree in [10], the errors brought by the proposed method are obviously less variable. Taking the SCPMA as an example, there are 2883 errors of "0" out of a total number of 3266 errors in Fig. 7 (e), which means more than 88.2% results remain the same. The error "2" in the same sub-plot shows the smallest number and looks odd, which is around 0.4% of all errors. In contrast, the SCPMA using SRP-PHAT in Fig. 7 (f) presents 3 errors of "2", 4 errors of "1" and 2 errors of "0" if rounded off to have the same resolution with Fig. 7 (e). This is nearly a uniform distribution, leading to random errors applying SRP-PHAT. For practical use, invariant errors are always preferable for the robustness of speech DOA estimation, which is an advantage of the proposed method over the SRP-PHAT approach in [10].

V. CONCLUSIONS

This paper explores the advantage of using CPMA and SCPMA for robustly estimating DOA of speech sources through a DL approach. In our previous work in [10], such advantages were not fully employed using SRP-PHAT, and the DOA estimates under high noise and strong reverberation were not satisfactory as expected. The proposed method alternatively designs a CNN-based network named CoprimeNet to train the GCC-PHAT extracted from speech recordings as a feature for DOA estimation. Existing research has shown the useful DOA information contained in GCC-PHAT, which is robust to noise and reverberation, and we improve the calculation of GCC-PHAT for SCPMA in this paper.

As an initial investigation in this direction for coprime array geometries, this paper evaluates the performance of speech DOA estimation under different levels of noise in a simulated large room with fixed source-array distance. The source is assumed far-field impinging from 55 degree to 125 degree. The speech corpus and configurations of CPMA, SCPMA and the comparative ULA with same numbers of microphones all remain the same with [10]. In the simulation, silence and unvoiced segments in speech recordings are removed before being recorded in the virtual room and mixed with white noise. GCC-PHAT is then extracted from the preprocessed recordings to be fed into the designed CoprimeNet, which progressively looks for patterns in the GCC-PHAT using a few blocks of convolutional layers and pooling layers. The CoprimeNet directly outputs DOA estimates by classifying information read from GCC-PHAT.

Experimental results show that the speech DOA estimation using ULA, CPMA and SCPMA are all accurate under different levels of noise, which present the benefit of applying DL to estimate speech DOA with coprime arrangements. The SCPMA performs the best in scenarios with 10 dB SNR and with no noise, meaning that the proposed GCC-PHAT formulated for SCPMA works well, and it is the SCPMA geometry that leads to more robust speech DOA estimation. Additional analysis on distributions of errors further demonstrates advantages of using DL over SRP-PHAT by bringing more stable DOA estimates, which is crucial for practical applications.

Future work will further investigate the advantage of using coprime-based microphone arrays and DL for robust speech DOA estimation. More complex acoustic environments will be trained and tested using CoprimeNet, including the combination of different levels of noise and reverberation. The dataset without truncation to recordings will also be simulated for training, which will contain silence and unvoiced segments. A broader range of DOAs will be tested, and for increasing the output resolution, the CoprimeNet may be remodelled to solve a regression task. In addition, we will look at improving the accuracy of speech DOA estimation using CPMA. The structure of CoprimeNet can also be improved to adapt to the aforementioned changes to data and experimental settings.

REFERENCES

[1] 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.

[2] J. M. Vera-Diaz, D. Pizarro, and J. Macias-Guarasa, "Towards End-to-End Acoustic Localization Using Deep Learning: From Audio Signals to Source Position Coordinates," *Sensors*, vol. 18, no. 10, Oct. 2018.

[3] T. Nakatani, R. Ikeshita, K. Kinoshita, H. Sawada and S. Araki, "Blind and Neural Network-Guided Convolutional Beamformer for Joint Denoising, Dereverberation and Source separation," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICCASP 2021)*, pp. 6129–6133, Jun. 2021.

[4] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "Robust Sound Source Tracking Using SRP-PHAT and 3D Convolutional Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 300–311, 2021.

[5] J. Benesty, J. Chen, and Y. Huang, *Microphone Arrays Signal Processing*, Springer-Verlag: Berlin, Germany, 2008.

[6] G. Huang, J. Benesty, and J. Chen, "Design of Robust Concentric Circular Differential Microphone Arrays," *The Journal of the Acoustical Society of America,* vol. 141, no. 5, pp. 3236–3249, May 2017.

[7] D. Bush and N. Xiang, "Broadband Implementation of Coprime Linear Microphone Arrays for Direction of Arrival Estimation," *The Journal of the Acoustical Society of America*, vol. 138, no. 1, pp. 447–456, Jul. 2015.

[8] J. Zhao and C. Ritz, "Investigating Co-Prime Microphone Arrays for Speech Direction of Arrival Estimation," *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2018)*, pp. 1658–1664, Nov. 2018.

[9] K. Adhikari, "Beamforming with Semi-Coprime Arrays," *The Journal of the Acoustical Society of America*, vol. 145, no. 5, pp. 2841–2850, May 2019.

[10] J. Zhao and C. Ritz, "Semi-Coprime Microphone Arrays for Estimating Direction of Arrival of Speech Sources", *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2019)*, Nov. 2019.

[11] S. Pasha, C. Ritz and Y. Zou, "Detecting Multiple, Simultaneous Talkers through Localising Speech Recorded by Ad-hoc Microphone Arrays," *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2016)*, Dec. 2016.

[12] J. Zhao and C. Ritz, "Co-Prime Circular Microphone Arrays and Their Application to Direction of Arrival Estimation of Speech Sources," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICCASP 2019)*, pp. 800–804, May 2019.

[13] J. H. DiBiase, "A High-accuracy, Low-latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays," *Brown University*, 2000.

[14] 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," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICCASP 2015),* pp. 2814–2818, Apr. 2015.

[15] Z. Liu, C. Zhang, and P. S. Yu, "Direction-of-Arrival Estimation Based on Deep Neural Networks with Robustness to Array Imperfections," *IEEE Transactions on Antennas and Propagation*, vol. 66, no. 12, Dec. 2018.

[16] S. Chakrabarty and E. A. P. Habets, "Multi-Speaker DOA Estimation Using Deep Convolutional Networks Trained with Noise Signals," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 8–21, Mar. 2019.

[17] E. L. Ferguson, S. B. Williams, and C. T. Jin, "Sound Source Localization in a Multipath Environment Using Convolutional Neural Networks," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICCASP 2018)*, pp. 2386–2390, Apr. 2018.

[18] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound Event Localization and Detection of Overlapping Sources Using Convolutional Recurrent Neural Networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, Mar. 2019.

[19] IEEE subcommittee on subjective measurements, "IEEE Recommended Practices for Speech Quality Measurements," *IEEE Transactions on Audio and Electroacoustics*, vol. 17, pp. 227–46, 1969.

[20] J. Zhao, C. Ritz, and J. Xi, "Investigating Co-prime Circular Microphone Arrays and Their Application to Speech Direction of Arrival Estimation," unpublished.

[21] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-scale Image Recognition," *International Conference on Learning Representations (ICLR 2015)*, May 2015.

[22] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *International Conference on Machine Learning (ICML 2015)*, Jul. 2015.

[23] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press: Cambridge, MA, USA, 2016.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Communications of the ACM*, vol. 60, no. 6, Jun. 2017.

[25] J. Allen and D. Berkley, "Image Method for Efficiently Simulating Small-room Acoustics," *The Journal of the Acoustical Society of America*, vol. 65, issue 4, pp. 943–950, Apr. 1979.