

Received 27 October 2023, accepted 27 November 2023, date of publication 5 December 2023,
date of current version 18 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3340108

RESEARCH ARTICLE

Time Delay Estimation for Sound Source Localization Using CNN-Based Multi-GCC Feature Fusion

HAITAO LIU^{1,2}, XIULIANG ZHANG¹, PENGGAO LI¹, YU YAO³, SHENG ZHANG⁴,
AND QIAN XIAO¹

¹School of Mechatronics and Vehicle Engineering, East China Jiaotong University, Nanchang 330013, China

²Suzhou Automotive Research Institute, Tsinghua University, Suzhou 215131, China

³School of Information and Telecommunications Engineering, Hainan University, Haikou 570288, China

⁴Suzhou Acoustic Technology Institute Company Ltd., Suzhou 518057, China

Corresponding authors: Haitao Liu (2860@ecjtu.edu.cn) and Yu Yao (996461@hainan.edu.cn)


This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 12104153; in part by the China Postdoctoral Science Foundation under Grant 2021M701963; and in part by the Training Plan for Academic and Technical Leaders of Major Disciplines in Jiangxi Province, China, under Grant 20204BCJL23034.

ABSTRACT Accurate time delay estimation is critical in sound source localization methods that rely on time difference of arrival. Background noise and reverberation often introduce errors in time delay estimation. Generalized cross-correlation (GCC) functions, paired with different weighting functions, can adapt to various sound field environments for time delay estimation. To create a highly accurate time delay estimation method suitable for universal sound field conditions, this paper proposes a novel approach, which involves training multi-class weighted generalized cross-correlation features using a convolutional neural network. Various weighted GCC functions are employed to extract time delay features for the same microphone pairs. These time delay features from multi-class weighted GCC are fused to create a feature matrix. The feature matrix is then input into a convolutional neural network composed of convolutional layers and fully connected layers for training and prediction. In the network, time delay estimation is achieved using two different methods: regression and classification, with mean squared error and cross-entropy serving as loss functions, respectively. The proposed method is tested and validated through simulation scenarios featuring various signal-to-noise ratios and reverberation conditions. Time delay estimation results are compared with recent state-of-the-art (SOTA) methods, assessing accuracy, root mean square error, and mean absolute error. The results demonstrate that the proposed method achieves an impressive 3.36% enhancement in overall delay estimation accuracy (within 10cm), reduces the absolute error by 11.53%, and significantly decreases the estimated root mean square error by 16.07% compared to existing SOTA methods. Furthermore, the proposed model offers the advantages of compact size and efficient computational performance when compared to existing methods. These findings underscore the exceptional comprehensive performance of the proposed model in sound source localization applications.

INDEX TERMS Sound source localization, time delay estimation, generalized cross-correlation, convolutional neural network, feature fusion.

I. INTRODUCTION

Time difference of arrival (TDOA) -based sound source localization algorithms is the most extensively studied

The associate editor coordinating the review of this manuscript and approving it for publication was Jon Atli Benediktsson .

two-step localization technique. It is evident that accurate time delay estimation (TDE) plays a crucial role in determining the positioning accuracy. And TDE plays a crucial role in various acoustic positioning applications, such as sound source tracking [1], robot localization [2], and self-calibration [3]. It serves as a fundamental component

in accurately determining the spatial information of the sound source in these scenarios. In a typical setup, TDE is performed by analyzing the signals received by a synchronized array of microphones with known positions. Since the transmitted waveform and its transmission time are unknown, direct determination of the time of travel from the sound source to the microphones is not possible. Instead, the TDE is measured by correlating the received signals. By collecting a set of TDE measurements from the microphone array, it becomes feasible to compute the direction of arrival (DOA) of the sound signal or even the position of the sound source using multilateration techniques [4].

The generalized cross-correlation (GCC) has been the most widely adopted method for TDOA estimation for many decades. In particular, the GCC with phase transform (GCC-PHAT) [5] filter is commonly used in many acoustic scenarios, due to its fast implementation and robustness in adverse environments. GCC-PHAT is based on a cross-correlation function between filtered versions of the received signals. The PHAT is a filter that uses the magnitude information of the cross-correlation to normalize the narrowband components, increasing the resolution of the TDOA function if compared to a simple cross-correlation. The GCC-PHAT is computed in the frequency domain using the fast Fourier transform (FFT) technique. This involves the calculation of the cross-spectrum, followed by the application of the PHAT filter. Finally, the inverse FFT is performed to obtain the time delay estimation function. The GCC-PHAT provides good TDE in moderate noisy and reverberation conditions. However, the estimation performance of the GCC-PHAT deteriorates significantly when reverberation or noise is high. Many methods have been proposed to improve robustness in adverse conditions. A class of TDE methods is based on the blind system identification [6], [7], [8], [9], which focuses on impulse responses between a source and the microphones. These methods require a certain time for the convergence of the filter to estimate the impulse responses, and in particular the direct path dominant peak. Thus, the practical application of this class of methods is very difficult. Other approaches exploit the use of redundant information among several microphones [10], [11], [12]. These methods are thus useful when more than a microphone pair is available.

With the recent advent of deep learning, a wide variety of methods for sound source localization have been developed. These methods process the raw waveforms or spectrograms of the signals without using cross-correlations [13]. And some of these methods train the neural network models to directly predict the DOA [14], [15] or sound source coordinates [16]. However, machine learning has also been applied in various cases to address the TDE problem. Wang et al. [17] proposed a cross-correlation with time-frequency masking predicted by a deep neural network based on bidirectional long short term memory networks. The goal of the deep learning masking is to emphasize the time-frequency units

dominated by the target speech. Ding et al. [18] processed the cross correlation sequence by a deep neural network with an output of 10 dimensional vector for TDE values. Comanducci et al. [19] proposed a frequency-sliding GCC with a convolutional neural network (CNN). The frequency-sliding allows the calculation of sub-band GCC for an arbitrary frequency band. The CNN serves as a fully convolutional denoising autoencoder, with its output representing the complete TDE function. Salvati et al. [20] introduced a scheme that employs a residual CNN to calculate TDE directly from the original waveform. This approach is utilized in the context of joint sound source recognition and localization tasks. In this approach, the PGCC-PHAT model is proposed, which entail the computation of parameterized GCC-PHAT and further processing of the outputs using a CNN to predict TDE for two signals. While this method effectively reduces the average error, it faces challenges when making precise predictions within a few samples, a critical aspect for achieving high-precision localization. Berg et al. [21] proposed the NGCC-PHAT model for TDE by filtering the raw waveforms using SincNet-based network before computing the GCC-PHAT. This network can be trained to exploit patterns in the data, e.g. the acoustic properties of human speech, in order to remove the effects of noise and reverberation. Furthermore, by using a shift equivariant CNN (SE-CNN), the network can learn to find useful representations while preserving the timing information contained in the signal. The method is able to consistently improve detection accuracy over the baseline GCC-PHAT and PGCC-PHAT. But the cost is that more calculations are required, which has a huge impact on the real-time performance of TDE. Shi et al. [22] introduced a dual-branch transformer CNN structure designed for face super-resolution. This approach leverages multiple neural networks to extract features, and the integration of features from different neural network branches is a noteworthy and valuable concept.

To enhance the performance of GCC for TDOA estimation in the presence of noise and reverberation, this paper presents a novel approach using CNN-Based Multi-GCC Feature fusion (MGCCFF) for TDE in sound source localization. The proposed method leverages multiple types of weighted generalized cross-correlation function (WGCCF), each with distinct delay estimation characteristics, to construct a feature matrix. By incorporating multiple WGCCFs, the feature matrix captures diverse time-delay information patterns derived from the same pair of microphone signals. This allows the model to exploit a rich set of temporal cues for more accurate delay estimation. The CNN is then employed to process the feature matrix, utilizing its ability to learn complex patterns and relationships in the data. By training the network on a large dataset, it can effectively extract discriminative features and estimate the time delay with improved accuracy. By fusing the multi-class WGCCF features using the CNN, the proposed method combines

TABLE 1. GCC weighting functions.

Name	$\varphi(k)$	Remark
Roth	$\frac{1}{G_{11}(k)}$	Effective suppression of noisy frequency bands.
SCOT	$\frac{1}{(G_{11}(k)G_{22}(k))}$	It has a certain elimination effect on the signal frequency with low signal-to-noise ratio.
Improved SCOT	$\frac{1}{(G_{11}(k)G_{22}(k))^{0.75}}$	False peak detections caused by ambient reverberation can be avoided.
PHAT	$\frac{1}{ G_{12}(k) }$	Strong ability to suppress noise.
Parameterized PHAT	$\frac{1}{ G_{12}(k) ^\beta}$	Estimation performance improvements for narrowband and wideband signals can be achieved when $\beta \in [0.5, 0.7]$ [28]
ML	$\frac{ \gamma(k) ^2}{ G_{12}(k) (1- \gamma(k) ^2)}$	Has a certain suppression effect on noise

In equation (5), $\tau \in [-\tau_{max}, \tau_{max}]$. τ_{max} is the maximal delay for a microphone pair, which is typically taken as

$$\tau_{max} = \|r_1 - r_2\| F_s / c \tag{6}$$

where c is the speed of sound and F_s is the sample rate. $\|r_1 - r_2\|$ represents the Euclidean distance between the microphone positions r_1 and r_2 . So the maximum time delay τ_{max} is measured in terms of the number of samples.

Theoretically, the significant peaks in Equation (5) represent sound sources, and the number of peaks is equal to the number of sound sources. The abscissa corresponding to the effective peak is the estimated value of TDE. The estimated time delay is then obtained as

$$\hat{\tau} = \text{argmax } R(\tau, k) \tag{7}$$

In the cross-correlation calculations, the introduction of a weighting function $\varphi(k)$ in the frequency domain is crucial. The reason is that reverberation and noise can significantly affect the estimation accuracy. Different weighting functions can help suppressing noise and reverberation, leading to more prominent peaks in the time delay calculation. As a result, various weighting functions with different sensitivities to different environments have been developed. One widely used example is the PHAT weighting function in the generalized cross-correlation algorithm, known for its robustness in adverse environments and its simplicity of implementation. Typical weighting functions from articles [5], [23], [24], [25], [26] are summarized in Table 1.

Supplement: $G_{11}(k)$ and $G_{22}(k)$ are the autopower spectrum functions of signals x_1 and x_2 respectively. $G_{12}(k)$ is the cross power spectral density function of signals x_1 and x_2 . Then the related expressions are as follows:

$$G_{11}(k) = X_1(k)X_1^*(k) \tag{8}$$

$$G_{22}(k) = X_2(k)X_2^*(k) \tag{9}$$

$$G_{12}(k) = X_1(k)X_2^*(k) \tag{10}$$

$$\gamma(k) = \frac{G_{12}(k)}{\sqrt{G_{11}(k)G_{22}(k)}} \tag{11}$$

Table 1 provides an overview of various weighting functions commonly used in TDE. The Roth weighting function, which is equivalent to Wiener filtering, effectively suppresses frequency bands with high noise but may result in a broader correlation function peak. The SCOT weighting function

incorporates the influence of noise and the interplay between two channels to emphasize coherent signal components while suppressing incoherent noise components, without necessarily leading to a discernible attenuation effect on signal frequencies with low signal-to-noise ratios. The improved SCOT weighting function addresses the issue of false peak detection caused by environmental reverberation and exhibits inhibitory effects on reverberation. The PHAT weighting function is widely employed and demonstrates strong noise suppression capabilities. The parameterized PHAT weighting function achieves improved estimation performance for both narrowband and wideband signals when the parameter β is within the range of [0.5, 0.7]. For this paper, a value of $\beta = 0.55$ is selected. The Maximum Likelihood (ML) weighting function employs an adaptive filter that assigns different weights to signals with varying signal-to-noise ratios, thereby effectively suppressing noise.

Different weighting functions exhibit distinct effects on environmental noise and reverberation. However, traditional methods based on theoretical mathematics often encounter challenges in achieving the optimal fusion of weighting function performance. In this paper, a novel approach is proposed to address this issue. The approach involves constructing a feature matrix by utilizing multiple weighting functions. Subsequently, targeted weighting and integration are performed through a neural network. The objective of this approach is to integrate an estimation model that effectively suppresses and eliminates environmental noise and reverberation.

C. DEFINITION OF FEATURE MATRIX

The key contribution of this method is the introduction of a feature matrix, which serves as the input to the CNN. The feature matrix is constructed by combining different weighted generalized cross-correlation functions. The feature matrix is defined as

$$\bar{R}(k) = \begin{bmatrix} R(\varphi_1, \tau_1, k) & R(\varphi_1, \tau_2, k) & \cdots & R(\varphi_1, \tau_d, k) \\ R(\varphi_2, \tau_1, k) & R(\varphi_2, \tau_2, k) & \cdots & R(\varphi_2, \tau_d, k) \\ \vdots & \vdots & \vdots & \vdots \\ R(\varphi_b, \tau_1, k) & R(\varphi_b, \tau_2, k) & \cdots & R(\varphi_b, \tau_d, k) \end{bmatrix} \tag{12}$$

where $\varphi_1, \varphi_2, \dots, \varphi_b$ represent the weighting functions in Table 1. $\tau_1 = -\tau_{max}, \tau_2 = -\tau_{max} + 1, \dots, \tau_d = \tau_{max}$, $d = 2\tau_{max} + 1$. The characteristic matrix $\bar{R}(k)$ is of size $b \times d$, and each row of the matrix is a generalized cross-correlation function with a specific weighting function. In theory, the corresponding τ values at the maximum value in each row are the TDE values estimated by the respective generalized cross-correlation functions.

D. DEFINING NONLINEAR MAPS: $F(\bullet, \alpha)$

To establish a mapping between the characteristic matrix $\bar{R}(k)$ of the signal received by the two microphones (with a length

of N) and the time delay information of the signal received by a single microphone, a neural network is employed to design a nonlinear mapping denoted as $F(\bullet, \alpha)$. Here, α represents parameters that is learned through the training process of the network.

The neural network architecture comprises several convolutional layers followed by fully connected layers. The TDE information is then obtained through the output layer. Data undergoes a filtering and activation detection step within the convolutional layer, as represented by the equation:

$$H^l = \sigma \left(W^l * H^{l-1} + b^l \right) \quad (13)$$

where, H^l and H^{l-1} denote feature maps in consecutive layers, W^l is a trained kernel, b^l is a bias parameter, $\sigma(\cdot)$ is the activation function, and $*$ represents convolution. The rectified linear unit (ReLU) is commonly used to generate the output of the convolutional layer. The bias ensures that each node has a trainable constant value. The output of the convolutional layers is then flattened to create a single feature vector, serving as the input for one or more fully connected layers. In a fully connected layer, each neuron is connected to all neurons of the preceding layer. A fully connected layer multiplies the input by a weight matrix and then adds a bias vector. The operation is defined as:

$$h_{fc}^l = \sigma \left(W_{fc}^l h_{fc}^{l-1} + b^l \right) \quad (14)$$

where h_{fc}^l represents the output of the fully connected layer at a specific layer l of the neural network. W_{fc}^l is the weight matrix associated with the fully connected layer at layer l . h_{fc}^{l-1} represents the input to the fully connected layer, which comes from the previous layer $l-1$. b^l denotes the bias vector associated with the fully connected layer at layer l . σ is the ReLU activation function applied element-wise to the linear transformation result.

In the proposed method, two different output methods, namely the regression-based methods and the regression-via-classification (RvC) methods, are designed as the output components of the nonlinear function $F(\cdot, \alpha)$.

The regression method tries to estimate $\bar{\tau}$ directly and the model is trained to minimize the mean squared error (MSE). For a single training example, the final predictions and MSE are defined as

$$\bar{\tau} = F \left(\bar{R}(k), \alpha \right) \quad (15)$$

$$MSE = \frac{1}{2} \sum_m \left(\bar{\tau} - \tau \right)^2 \quad (16)$$

where $\bar{\tau}$ represents the output of the network, which corresponds to the estimated time delay of MGCCFF model. τ represents the correct time delay, m is the input sample size. $\bar{R}(k)$ is the feature matrix.

The RvC method utilizes softmax normalization in the last layer to generate a probability distribution over time delays.

This approach aims to consolidate information from various correlations and transform it into a probability distribution that reflects the likelihood of a sound source occurring at each specific time delay. By applying softmax normalization, the probabilities are constrained to sum up to 1, enabling a meaningful representation of the distribution. Consequently, the final predictions are obtained as

$$P(\tau|x_1, x_2) = F \left(\bar{R}(k), \alpha \right) \quad (17)$$

$$\bar{\tau} = \operatorname{argmax} P(\tau|x_1, x_2) \quad (18)$$

where $P(\tau|x_1, x_2)$ contains the probabilities for each time delay $\bar{\tau} = -\tau_{\max}, \dots, \tau_{\max}$ considered in the correlation.

The networks $F(\bullet, \alpha)$ can be trained by minimizing the cross-entropy (CE) loss function, which for a single training example becomes:

$$CE = - \sum_m t_m \operatorname{Log} (P(\tau_m|x_1, x_2)) \quad (19)$$

where \sum_m is summation over all possible values of time delay. t_m is the true label value, encoded using one-hot encoding where only one element is 1, indicating the correct time delay, and others are 0. $P(\tau_m|x_1, x_2)$ is the predicted probability assigned by the model to the time delay τ_m given inputs x_1 and x_2 . The equation calculates the logarithm of the predicted probability for the correct τ and multiplies it by the corresponding true label value. This is done for all possible τ values, and the results are summed together, yielding the overall Cross-Entropy Loss. The objective of this loss function is to minimize the discrepancy between the predicted probabilities and the true labels, thereby enhancing the accuracy of τ prediction.

E. CNN ARCHITECTURE

In this study, a valid frame length of 2048 samples (128 ms) is used with a sampling rate of 16 kHz. The distance between the microphones is 0.5 m, and the speed of sound is assumed to be 343 m/s. Based on these parameters, the theoretical maximum delay represented by the number of sampling points can be calculated by equation (6), which is 23. Since the value range of τ in GCC calculation is from $-\tau_{\max}$ to τ_{\max} , inclusive, the number of columns in the feature matrix can be determined as: $d = 2\tau_{\max} + 1 = 47$. This means that the feature matrix will have 47 columns, and each corresponding to a specific time delay value within the range of $-\tau_{\max}$ to τ_{\max} . For six types of WGCCFs, the dimensions of resulting feature matrix is $b \times d = 6 \times 47$, namely 6 rows and 47 columns. Considering that the convolution operation of the first layer reduces the number of columns of the feature matrix by 2, the range of possible delays is reduced. In order to avoid this situation, when taking the WGCCF value to form the feature matrix, this paper reasonably takes two more values corresponding to the nearest delay range ($p=2$). Therefore, the final feature matrix is $b \times (d + 2) = 6 \times 49$, shown as Fig.3.

The architecture of the CNN in this study consists of 2 two-dimensional convolutional layers and 3 fully connected layers. After each convolutional layer, batch normalization is applied. In the first convolutional layer, there are 32 filters, and this number is doubled for the subsequent convolutional layer. The kernel size for the first convolutional layer is 1×3 for TDE feature extraction, while for the second convolutional layer, it is 6×1 TDE feature extraction. To enhance nonlinearity and reduce overfitting, three fully connected layers are used, with two dropout layers inserted between them. The dropout layer has a probability of 0.2. The first and second fully connected layers have 64 neurons each. The last fully connected layer has 1 or 47 neurons for regression output or classification output respectively. Fig.3 illustrates the architecture of the CNN for the RvC and regression output.

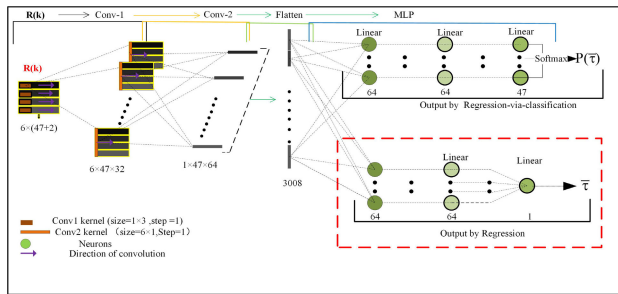


FIGURE 3. CNN architecture.

III. SIMULATIONS

In order to evaluate the performance of the method proposed in this paper, the flow chart of the simulation experiment shown as Fig.4

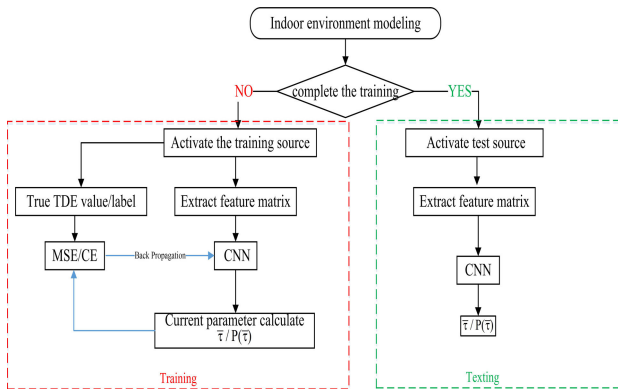


FIGURE 4. The flow chart of the simulation experiment.

A. INDOOR ENVIRONMENT MODELING

To simulate realistic sound propagation and capture the effects of reverberation in indoor environments, the Pyroomacoustics [27] library is utilized. Pyroomacoustics offers a comprehensive set of tools and algorithms for modeling and simulating acoustic environments based on the image source method [28].

B. DATASET SELECTION AND PREPROCESSING

The audio signals were collected from the LibriSpeech dataset [29], which contains speech recordings from read audiobooks in English, sampled at $F_s = 16$ kHz. The data split based on speakers, such that 40 speakers were used for training, 3 for validation and 3 for testing. For each recording, silent parts removed by using a voice activity detector and then a 2 second long snippet from each recording extracted. This results in 1892 snippets for training (corresponding to roughly one hour of audio), 188 for validation and 216 for testing. During training, a frame of $N = 2048$ samples is randomly sampled for each snippet, while each of 15 non-overlapping windows are evaluated during testing, for a total of $216 * 15 = 3240$ time delay estimates.

By using the LibriSpeech dataset and following the described data processing steps, the study obtained a diverse and representative set of audio snippets for training, validation, and testing, allowing for robust evaluation of the proposed method.

C. TRAINING NETWORK

Network training was done inside a simulated room of dimension $7 * 5 * 3$ m, with microphones placed roughly in the middle of the room at $r_1 = [3.5, 2.25, 1.5]^T$ m and $r_2 = [3.5, 2.75, 1.5]^T$ m from the origin. This setup results in a maximum delay of $\tau_{max} = 23$ samples. The source positions r_s were sampled randomly from a uniform distribution over the entire room for each training sample. Furthermore, random reverberation times T_{60} and SNR were sampled in the ranges $[0.2, 1.0]$ s and $[0, 30]$ dB respectively. The Adam optimizer is used with a batch size of 32, a learning rate of 0.001 with a cosine decay schedule and train the network for 50 epochs.

By training the network with data generated in the simulated room environment, the model learns to effectively handle various room configurations, reverberation times, and signal-to-noise ratios, improving its generalization capabilities for real-world scenarios.

D. TESTING NETWORK

The trained models were evaluated in a different room with dimensions $6 * 4 * 2.5$ m and with the microphones placed at $r_1 = [3, 1.75, 1.25]^T$ and $r_2 = [3, 2.25, 1.25]^T$ m respectively, and the source positions were again sampled randomly across the whole room. Each recording was evaluated for $SNR \in [0, 6, 12, 18, 24, 30]$ dB and reverberation times $T_{60} \in [0.2, 0.4, 0.6, 0.8, 1]$ s.

By evaluating the trained models in this different room environment with varying SNR and reverberation conditions, the performance and generalization capabilities of the models can be assessed in more realistic and diverse scenarios.

E. COMPARISON MODEL

In order to provide a comprehensive comparison, the GCC-PHAT, PGCC-PHAT, and NGCC-PHAT methods were implemented and evaluated alongside the proposed method

in this paper. The GCC-PHAT is a widely used TDE method that calculates the cross-correlation between two microphone signals after applying the phase transform. The PGCC-PHAT method extends the GCC-PHAT method by using a CNN to predict the time delay. It takes multiple differently weighted GCC-PHAT as input and combines them into a single time delay prediction. Each correlation has a different weighting filter parameter β in the range of $[0, 0.1, \dots, 1]$. The network is trained to minimize the MSE loss. The NGCC-PHAT method utilizes a SincNet network to filter the received signals, applies the GCC-PHAT method to calculate the delay for each filtered signal, and then uses a neural network for probability estimation of each possible TDE value through classification. The network is trained to minimize CE loss. According to our research, the NGCC-PHAT method introduced by Berg et al. [21] represents the recent SOTA method.

In the subsequent experimental study of the mixing feature, the paper demonstrates that the MGCCFF method uses different loss functions (MSE for regression, CE for RvC) to train the model, emphasizing different evaluation metrics. To ensure a fair comparison, the MGCCFF method in this paper employs the same loss functions as the comparison models during training and evaluation. This allows for a consistent evaluation of the accuracy, root mean square error (RMSE), and mean absolute error (MAE) metrics for TDE. In comparison to the PGCC-PHAT model, this paper evaluates the accuracy, root mean square error, and absolute error of the estimated time delay of the MGCCFF model using mean square error as the loss function (MGCCFF-MSE). And, in comparison to the NGCC-PHAT model, this paper evaluates the accuracy, mean square error, and absolute error metrics for TDE of the MGCCFF model using cross entropy as the loss function (MGCCFF-CE).

IV. SIMULATION RESULTS AND ANALYSIS

To evaluate the performance of the proposed MGCCFF model, a series of group experiments were conducted in this paper. And all models underwent training, testing, and evaluation within the same simulation environment using the same dataset. This standardized approach ensures a fair comparison among the models and allows for a reliable assessment of their performance. By utilizing a consistent simulation environment and dataset, any discrepancies in results can be attributed to the differences in model architectures, input features, or other experimental factors, rather than variations in the experimental setup. The experiments were conducted to evaluate the MAE, RMSE, and Acc of each model under different SNR and reverberation times. It should be noted that all estimated delays, initially represented in terms of the number of samples, have been converted to their corresponding delays in millisecond. This conversion allows for a more meaningful interpretation and analysis of the results in real-world units of measurement. To assess accuracy, the metric used is the probability $P(|\bar{\tau} - \tau| < \eta/c)$, where η represents a threshold distance error of 10 cm.

This threshold is commonly employed to gauge the average accuracy achievable by acoustic localization systems [30], [31]. The threshold provides a criterion for assessing the precision of the model's TDE. By analyzing the percentage of accurate estimations within the specified level of precision, one can evaluate the performance and effectiveness of the model in estimating time delays.

A. FEATURE FUSION EXPERIMENT

The "Feature Fusion Experiment" depicted in Fig. 5 is designed to investigate the impact of the network model, the number of fused features, and the choice of model loss function on the overall performance of the model.

Fig. 5 presents a comparison of the chosen fused features in four different cases: one type, two types, four types, and six types. These cases refer to the selection and combination of different types of features for fusion. In the "1-MSE(CE)" group experiment, a single GCC-PHAT was chosen as the feature matrix training model. In the "2-MSE(CE)" group experiment, the feature matrix was constructed using both GCC-PHAT and GCC-ML. In the "4-MSE(CE)" group experiment, the feature matrix was constructed using GCC-Roth, GCC-ML, GCC-SCOT, and GCC-PHAT. Finally, in the "6-MSE(CE)" group experiment, all available weight-GCC functions were used to construct the feature matrix. For each group, the suffix "MSE" indicates that the training model minimizes the root mean square error and utilizes regression output as the output method. Conversely, the suffix "CE" indicates that the model is trained to minimize cross-entropy and utilizes regression-via-classification as the output method.

From Fig. 5, the following conclusions can be drawn:

- 1) The comparison between the GCC-PHAT model and the 1-MSE(CE) model demonstrates that the neural network model proposed in this paper is capable of continuously improving estimation accuracy and reducing estimation errors.
- 2) Comparing models with the same neural network architecture, but different input features (represented by the same line type in the Fig.5), there is convincing evidence that the mixed feature model outperforms the single GCC-PHAT feature model in terms of estimation accuracy and robustness. And the more features fused, the better the estimation accuracy and the lower the estimation error. These findings underscore the importance of combining different features in models, as it captures various aspects of the data and provides more comprehensive information for precise estimation.
- 3) Comparing models with the same input features but different loss functions (represented by the same line color model in Fig. 5), it becomes apparent that the cross-entropy loss function is more effective in enhancing the Acc of the model, while the mean square error loss function leads to smaller RMSE values.

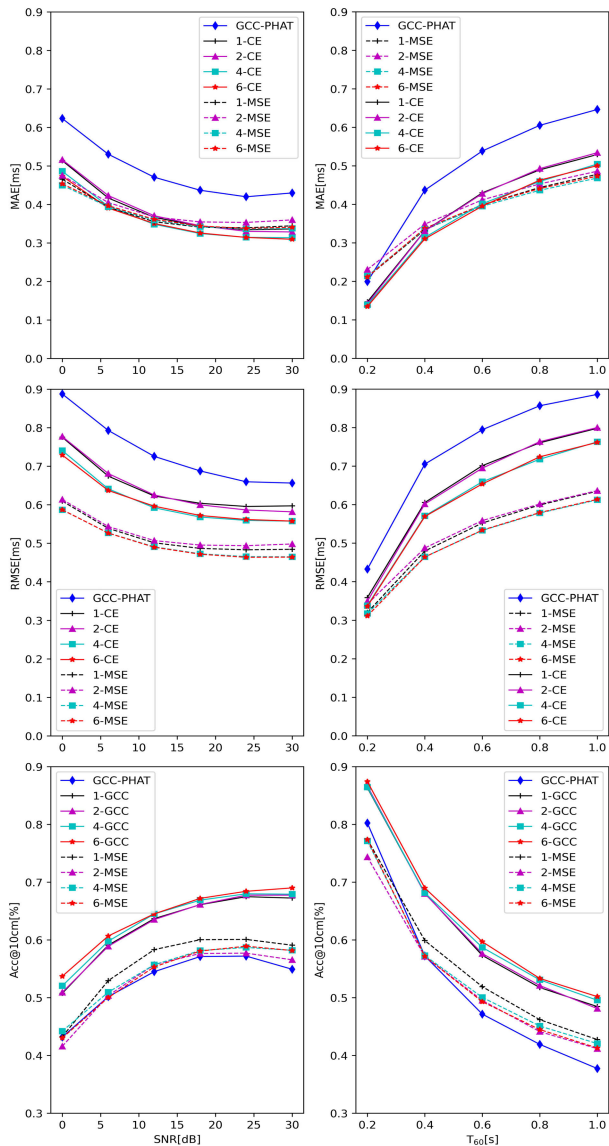


FIGURE 5. RMSE/Acc/MAE for different SNR and reverberation time.

B. MGCCFF-MSE MODEL VS. PGCC-PHAT MODEL

The PGCC-PHAT model, which employs the MSE as its loss function, has been established as a robust method that effectively minimizes the RMSE for delay estimation. In this paper, the proposed MGCCFF method is evaluated by comparing its performance in terms of Acc, RMSE, and MAE with the PGCC-PHAT model, using the same MSE loss function(MGCCFF-MSE).

The results in Fig. 6 to 10 clearly demonstrate that the robustness of the MGCCFF-MSE method consistently outperforms the GCC-PHAT and PGCC-PHAT methods across different signal-to-noise ratios and reverberation times. It achieves the lowest values for both RMSE and MAE for the overall estimate, indicating superior accuracy (within 10cm) in time delay estimation. Furthermore, it can be seen from Fig.9 that the proposed method exhibits exceptional performance in challenging scenarios characterized by high

reverberation ($T_{60} > 0.4$ s) and high SNR ($SNR > 7$ dB). In these conditions, the proposed method achieves the highest level of accuracy among all the evaluated models. However, in the case of low echo, the model exhibits lower estimation accuracy and higher Mean Absolute Error (MAE) compared to GCC-PHAT. This discrepancy can be attributed to potential overfitting during the model training process. A comparison of error distributions for the different methods in a high SNR environment can be seen in Fig.10. Because both the PGCC-PHAT and MGCCFF-MSE models are trained to minimize the MSE, their error distributions tend to have smaller tails compared to GCC-PHAT. Additionally, the MGCCFF-MSE model exhibits even smaller tails in its error distribution. This observation further demonstrates the effectiveness of the MGCCFF model, which leverages Multi-class GCC features, in improving the overall performance and reducing the occurrence of large errors.

In summary, the MGCCFF method with MSE loss function consistently achieves the lowest RMSE values across various conditions, surpassing the performance of the GCC-PHAT and PGCC-PHAT methods. Additionally, it excels in accurately estimating time delays in high reverberation and high signal-to-noise ratio environments.

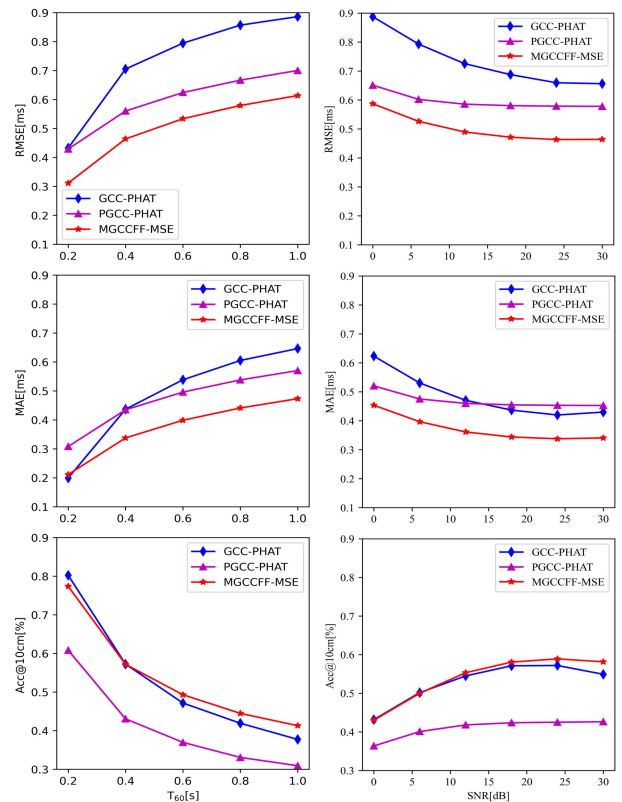


FIGURE 6. RMSE/MAE/Acc for different SNR and reverberation time.

C. MGCCFF-CE MODEL VS. NGCC-PHAT MODEL

The NGCC-PHAT model uses cross-entropy(CE) as the loss function, which can continuously improve the model estimation accuracy and reduce the absolute error. In this

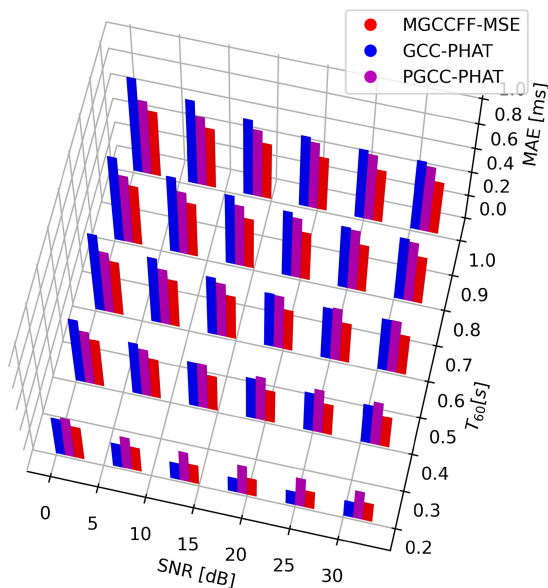


FIGURE 7. MAE under each set of SNR and T_{60} .

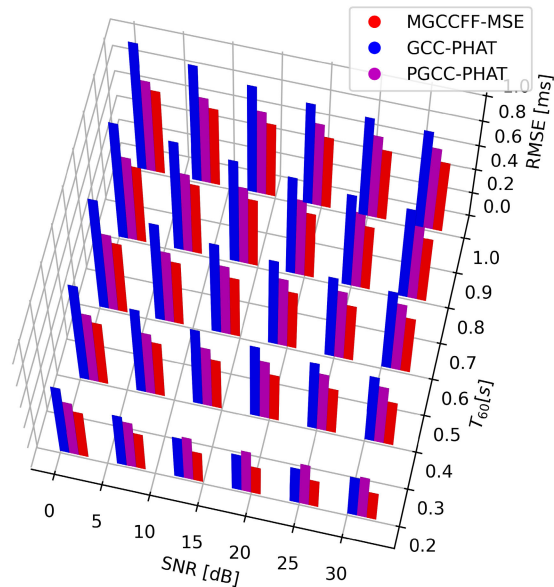


FIGURE 8. RMSE under each set of SNR and T_{60} .

paper, the proposed MGCCFF method is evaluated by comparing its performance in terms of Acc, RMSE, and MAE with the PGCC-PHAT model, using the same CE loss function(MGCCFF-CE).

Based on the results presented in Fig. 11 to 15, the following conclusions can be drawn from this study. The MGCCFF-CE method proposed in this paper consistently achieves the highest localization accuracy across various SNR and reverberation levels. Additionally, the proposed method exhibits lower absolute error in time delay estimation. In Fig. 15, the error distributions of various methods are compared in an environment with a reverberation time of $T_{60} = 0.4$ s and a SNR of 30 dB. It is observed that the MGCCFF-CE model exhibits fewer incorrect predictions at 0 delay compared to the other methods. This implies that the

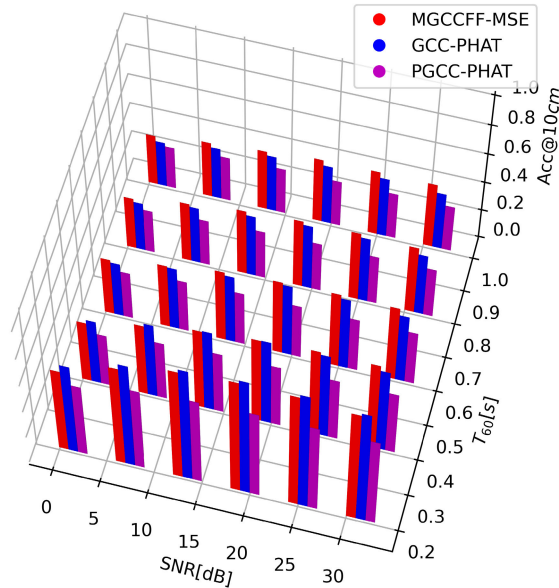


FIGURE 9. Acc under each set of SNR and T_{60} .

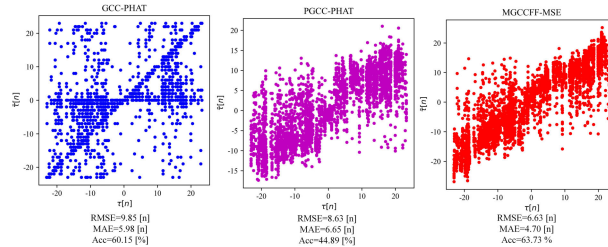


FIGURE 10. Scatter plots of ground truth and predicted time delays for reverberation time $T_{60} = 0.4$ s and SNR = 30 dB.

MGCCFF-CE model performs better in accurately estimating the time delay under these specific conditions.

However, in scenarios with high signal-to-noise ratio and low reverberation, the performance of the MGCCFF-CE model is not exceptional. Specifically, the estimated RMSE and MAE are a little larger compared to the NGCC-PHAT model but consistently smaller than those of the GCC-PHAT model. This performance difference can be attributed to the unique advantages of the SincNet network employed by the NGCC-PHAT model. In contrast, the MGCCFF model directly extracts GCC features from the original signal, simplifying the processing pipeline. While this approach offers simplicity, it may result in a limited set of features compared to the Sincnet network. This limitation becomes more pronounced in scenes with high SNR and low reverberation. This is due to the fact that the application of SincNet preprocessing enhances the quality of the signal by reducing noise and improving the representation of the underlying signal. Consequently, the improved signal quality results in enhanced performance, particularly in terms of higher SNR and lower T_{60} values.

D. DETAILED COMPARISON OF SELECTED MODELS

Table 2 presents a comprehensive comparison of the models discussed. In the PGCC-PHAT model, the feature matrix

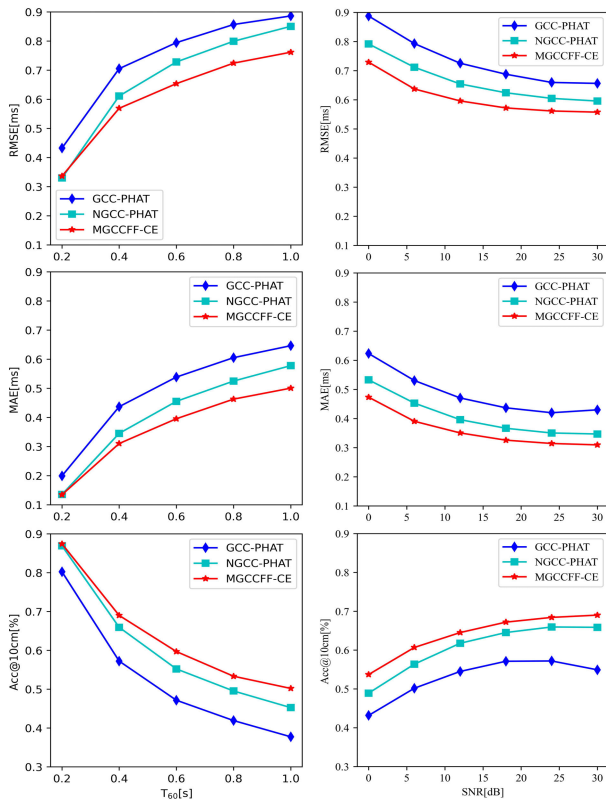


FIGURE 11. RMSE/MAE/Acc for different SNR and reverberation time.

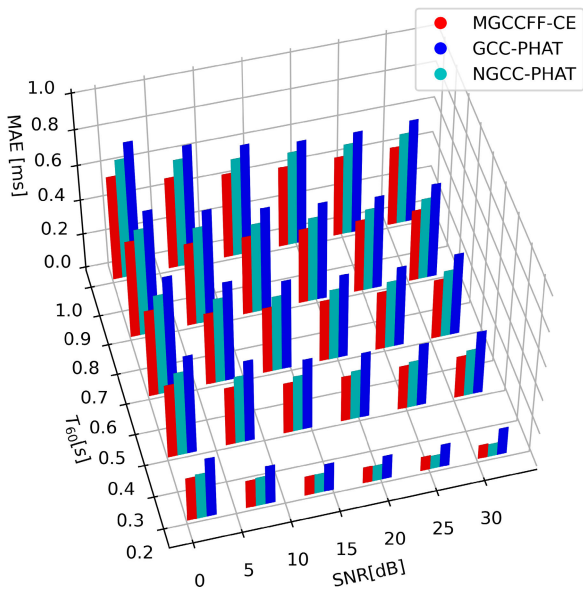


FIGURE 12. MAE under each set of SNR and T_{60} .

is constructed using the parameterized phase transformation weighted generalized cross-correlation function. The model employs a 3×3 convolution kernel and consists of 5 convolutional layers. The number of convolution kernels in each layer is twice that of the previous layer, starting with 32 kernels in the first layer. This design choice results in a more complex model with a larger number of parameters.

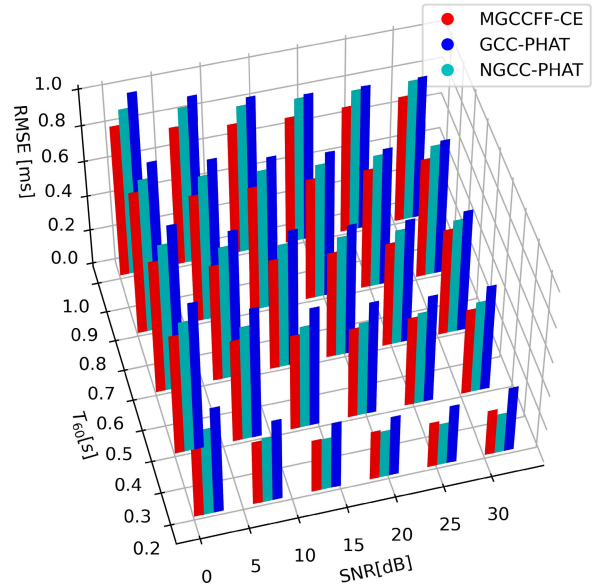


FIGURE 13. RMSE under each set of SNR and T_{60} .

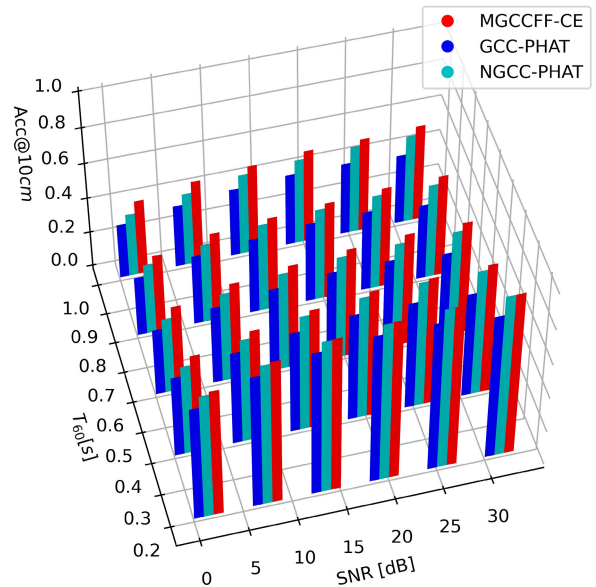


FIGURE 14. Acc under each set of SNR and T_{60} .

The NGCC-PHAT model incorporates the Sincnet model in the first convolutional layer to filter the original signal. This process creates 128 independent filtering channels, resulting in 128 different versions of the filtered signal. Parallel convolution calculations and classification are then performed. Due to this additional complexity and the need for more extensive calculations and memory space, the NGCC-PHAT model is relatively complex. In contrast, the MGCCFF method proposed in this paper adopts a simpler approach. It utilizes only 6 different weighting functions and incorporates two convolutional layers with smaller convolution kernels (1×3 and 6×1 , respectively). Consequently, the model is relatively lightweight with a reduced number of parameters.

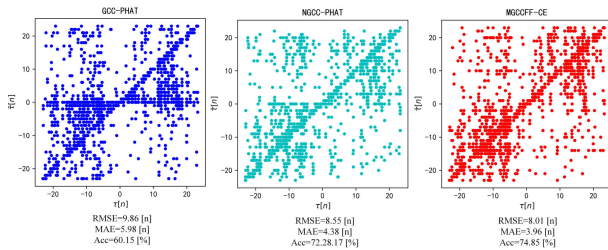


FIGURE 15. Scatter plots of ground truth and predicted time delays for reverberation time $T_{60} = 0.4s$ and $SNR = 30dB$.

TABLE 2. Detailed comparison of models.

Model	RMSE (ms)	MAE (ms)	Acc (%)	#params	FLOPs	Size
GCC-PHAT	0.7351	0.4853	0.5285	-	-	-
PGCC-PHAT	0.5962	0.4696	0.4097	109MB	46.14MB	47.89MB
NGCC-PHAT	0.6638	0.4078	0.6057	2.37G	4.08MB	37.95MB
MGCCFF-MSE	0.5004	0.3725	0.5393	0.81MB	0.84MB	1.05MB
MGCCFF-CE	0.6090	0.3608	0.6393	0.82MB	0.85MB	1.06MB

In summary, the PGCC-PHAT, NGCC-PHAT and MGCCFF models differ in terms of complexity, parameterization, and computational requirements. The PGCC-PHAT model is the most complex, while the method proposed in this paper offers a simpler alternative with fewer parameters.

The table results demonstrate that the MGCCFF-MSE method proposed in this paper exhibits optimal robustness and improved estimation accuracy. The overall RMSE for time delay estimation is approximately 31.93% lower than that of the traditional GCC-PHAT method, and about 16.07% lower than the PGCC-PHAT method. Furthermore, the MGCCFF-CE model in this paper demonstrates the best accuracy and good robustness in TDE. The overall estimation accuracy (within 10cm) is approximately 11.08% higher than that of GCC-PHAT and about 3.36% higher than NGCC-PHAT. Additionally, the average absolute error in overall time delay estimation is the lowest, with a reduction of approximately 25.65% compared to GCC-PHAT, and about 11.53% compared to NGCC-PHAT.

Moreover, the computational requirements of the method proposed in this paper are significantly lower than the other two models. The MGCCFF-MSE (CE) model in this paper only requires approximately 0.81 (0.82) MB of calculation, whereas the PGCC-PHAT and NGCC-PHAT demand 109MB and 2.39GB respectively. This indicates that the proposed method has a smaller computational load and a simpler model structure. Additionally, the memory footprint of the proposed model is only approximately 1.05 (1.06) MB, while the PGCC-PHAT and NGCC-PHAT occupy 47.89 MB and 37.95 MB respectively. Therefore, the model proposed provides significant advantages for engineering applications.

V. CONCLUSION

This paper presents a novel method for time delay estimation in the field of sound source localization with a microphone array. The proposed method leverages a fusion

feature matrix obtained by combining multiple weighted generalized cross-correlation functions with distinct characteristics. The time difference between the sound source and microphone pairs is then estimated using a CNN. By employing MGCCFF, the proposed method achieves substantial improvements in the accuracy, root mean square error, and average absolute error levels of time delay estimation. The experimental results demonstrate the excellent overall performance of the proposed model, highlighting its effectiveness in accurately estimating time delays in sound source localization tasks.

For the proposed MGCCFF model, the MGCCFF-CE method significantly improves estimation accuracy and reduces estimation error compared to the GCC-PHAT and NGCC-PHAT methods, and exhibits better robustness and achieves optimal accuracy in time delay estimation. On the other hand, the MGCCFF-MSE method effectively reduces the error in time delay estimation in indoor environments compared to the GCC-PHAT and PGCC-PHAT methods. The proposed method demonstrates superior robustness and maintains good detection accuracy across various levels of noise ratio and reverberation. Meanwhile, the MGCCFF model is able to achieve accurate results while minimizing computational resource requirements, providing a promising solution for real-world implementation and deployment.

In future work, we aim to enhance the performance of MGCCFF model by integrating cutting-edge deep learning concepts and implementing suitable preprocessing techniques on the original signal.

REFERENCES

- [1] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "Robust sound source tracking using SRP-PHAT and 3D convolutional neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 300–311, 2021.
- [2] X. Li, L. Girin, F. Badeig, and R. Horaud, "Reverberant sound localization with a robot head based on direct-path relative transfer function," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2016, pp. 2819–2826.
- [3] S. Burgess, Y. Kuang, and K. Åström, "TOA sensor network self-calibration for receiver and transmitter spaces with difference in dimension," *Signal Process.*, vol. 107, pp. 33–42, Feb. 2015.
- [4] K. Åström, M. Larsson, G. Flood, and M. Oskarsson, "Extension of time-difference-of-arrival self calibration solutions using robust multilateration," in *Proc. 29th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2021, pp. 870–874.
- [5] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, no. 4, pp. 320–327, Aug. 1976.
- [6] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *J. Acoust. Soc. Amer.*, vol. 107, no. 1, pp. 384–391, Jan. 2000.
- [7] S. Doclo and M. Moonen, "Robust adaptive time delay estimation for speaker localization in noisy and reverberant acoustic environments," *EURASIP J. Adv. Signal Process.*, vol. 2003, no. 11, pp. 1–15, Dec. 2003.
- [8] J.-W. Cho and H.-M. Park, "Imposition of sparse priors in adaptive time delay estimation for speaker localization in reverberant environments," *IEEE Signal Process. Lett.*, vol. 16, no. 3, pp. 180–183, Mar. 2009.
- [9] D. Salvati and S. Canazza, "Adaptive time delay estimation using filter length constraints for source localization in reverberant acoustic environments," *IEEE Signal Process. Lett.*, vol. 20, no. 5, pp. 507–510, May 2013.
- [10] J. Chen, J. Benesty, and Y. Huang, "Robust time delay estimation exploiting redundancy among multiple microphones," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 549–557, Nov. 2003.

- [11] J. Scheuing and B. Yang, "Disambiguation of TDOA estimation for multiple sources in reverberant environments," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 8, pp. 1479–1489, Nov. 2008.
- [12] H. Sundar, T. V. Sreenivas, and C. S. Seelamantula, "TDOA-based multiple acoustic source localization without association ambiguity," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 11, pp. 1976–1990, Nov. 2018.
- [13] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A survey of sound source localization with deep learning methods," *J. Acoust. Soc. Amer.*, vol. 152, no. 1, pp. 107–151, Jul. 2022.
- [14] S. Chakrabarty and E. A. P. Habets, "Broadband DOA estimation using convolutional neural networks trained with noise signals," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, Oct. 2017, pp. 136–140.
- [15] T. N. T. Nguyen, N. K. Nguyen, H. Phan, L. Pham, K. Ooi, D. L. Jones, and W.-S. Gan, "A general network architecture for sound event localization and detection using transfer learning and recurrent neural network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 935–939.
- [16] J. 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, p. 3418, Oct. 2018.
- [17] Z.-Q. Wang, X. Zhang, and D. Wang, "Robust TDOA estimation based on time-frequency masking and deep neural networks," in *Proc. Interspeech*, Sep. 2018, pp. 322–326.
- [18] J. Ding, B. Ren, and N. Zheng, "Microphone array acoustic source localization system based on deep learning," in *Proc. 11th Int. Symp. Chin. Spoken Language Process. (ISCSLP)*, Nov. 2018, pp. 409–413.
- [19] L. Comanducci, M. Cobos, F. Antonacci, and A. Sarti, "Time difference of arrival estimation from frequency-sliding generalized cross-correlations using convolutional neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 4945–4949.
- [20] D. Salvati, C. Drioli, and G. L. Foresti, "Two-microphone end-to-end speaker joint identification and localization via convolutional neural networks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–6.
- [21] A. Berg, M. O'Connor, K. Åström, and M. Oskarsson, "Extending GCC-PHAT using shift equivariant neural networks," 2022, *arXiv:2208.04654*.
- [22] J. Shi, Y. Wang, Z. Yu, G. Li, X. Hong, F. Wang, and Y. Gong, "Exploiting multi-scale parallel self-attention and local variation via dual-branch transformer-CNN structure for face super-resolution," *IEEE Trans. Multimedia*, early access, pp. 1–14, 2023.
- [23] G. C. Carter, A. H. Nuttall, and P. G. Cable, "The smoothed coherence transform," *Proc. IEEE*, vol. 61, no. 10, pp. 1497–1498, Mar. 1973.
- [24] P. R. Roth, "Effective measurements using digital signal analysis," *IEEE Spectr.*, vol. S-8, no. 4, pp. 62–70, Apr. 1971.
- [25] E. J. Hannan and P. J. Thomson, "Estimating group delay," *Biometrika*, vol. 60, no. 2, pp. 241–253, 1973.
- [26] K. D. Donohue, J. Hannemann, and H. G. Dietz, "Performance of phase transform for detecting sound sources with microphone arrays in reverberant and noisy environments," *Signal Process.*, vol. 87, no. 7, pp. 1677–1691, Jul. 2007.
- [27] R. Scheibler, E. Bezzam, and I. Dokmanic, "Pyroomacoustics: A Python package for audio room simulation and array processing algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 351–355.
- [28] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.
- [29] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 5206–5210.
- [30] L. Wang, N. Kitaoka, and S. Nakagawa, "Robust distant speaker recognition based on position-dependent CMN by combining speaker-specific GMM with speaker-adapted HMM," *Speech Commun.*, vol. 49, no. 6, pp. 501–513, Jun. 2007.
- [31] A. Plinge, F. Jacob, R. Haeb-Umbach, and G. A. Fink, "Acoustic microphone geometry calibration: An overview and experimental evaluation of state-of-the-art algorithms," *IEEE Signal Process. Mag.*, vol. 33, no. 4, pp. 14–29, Jul. 2016.



HAITAO LIU received the B.E. degree in vehicle engineering from Jilin University, Changchun, China, in 2009, and the Ph.D. degree in vehicle engineering from Tsinghua University, China, in 2015. From 2018 to 2019, he was a Visiting Scholar with the Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON, Canada. He is currently an Associate Professor with the School of Mechatronics and Vehicle Engineering, East China Jiaotong University, Nanchang, China. His current research interests include array signal processing, information fusion, sound source identification and localization, sound source tracking, vibration, and noise control.



XIULIANG ZHANG is currently pursuing the M.S. degree with the School of Mechatronics and Vehicle Engineering, East China Jiaotong University. His current research interests include array signal processing, sound source localization, reverberation elimination, and noise reduction.



PENGGAO LI is currently pursuing the M.S. degree with the School of Mechatronics and Vehicle Engineering, East China Jiaotong University. His current research interests include array signal processing and sound source localization.



YU YAO was born in 1986. He received the B.S. degree from Nanchang Hangkong University, Nanchang, China, in 2007, and the M.S. and Ph.D. degrees from Southeast University, Nanjing, China, in 2015. He is currently an Associate Professor with the School of Information and Telecommunications Engineering, Hainan University. His research interests include the communication and radar signal processing and cognitive radio, particularly with the radar-communication integration.



includes acoustic and related signal processing.

SHENG ZHANG received the B.S. and M.S. degrees from the Nanjing University of Aeronautics and Astronautics, in 2013 and 2016, respectively. He was an Acoustic Engineer with Knowles Electronics for two years. After that, he was with the Suzhou Automotive Research Institute, Tsinghua University, as a Senior Software Engineer for three years. He is currently the Technical Director of Suzhou Acoustic Technology Institute Company Ltd. His research interest



QIAN XIAO received the Ph.D. degree in vehicle engineering from the China Academy of Railway Sciences, China. He is currently a Professor with the School of Mechatronics and Vehicle Engineering, East China Jiaotong University, Nanchang, China. His research interests include track vehicle wheeltrack relation, railway vehicle running quality analysis and evaluation, and CAD/CAM/CAE.