applied sciences

MDPI

Article

A Deep Learning Method for DOA Estimation with Covariance Matrices in Reverberant Environments

Qinghua Huang and Weilun Fang *

Key Laboratory of Specialty Fiber Optics and Optical Access Networks, Shanghai University, Shanghai 200444, China; qinghua@shu.edu.cn
* Correspondence: fangweilun@shu.edu.cn

Abstract: Acoustic source localization in the spherical harmonic domain with reverberation has hitherto not been extensively investigated. Moreover, deep learning frameworks have been utilized to estimate the direction-of-arrival (DOA) with spherical microphone arrays under environments with reverberation and noise for low computational complexity and high accuracy. This paper proposes three different covariance matrices as the input features and two different learning strategies for the DOA task. There is a progressive relationship among the three covariance matrices. The second matrix can be obtained by processing the first matrix and it effectively filters out the effects of the microphone array and mode strength to some extent. The third matrix can be obtained by processing the second matrix and it further efficiently removes information irrelevant to location information. In terms of the strategies, the first strategy is a regular learning strategy, while the second strategy is to split the task into three parts to be performed in parallel. Experiments were conducted both on the simulated and real datasets to show that the proposed method has higher accuracy than the conventional methods and lower computational complexity. Thus, the proposed method can effectively resist reverberation and noise.

Keywords: direction-of-arrival; spherical microphone array; covariance matrix; convolutional neural network

Citation: Huang, Q.; Fang, W. A Deep Learning Method for DOA Estimation with Covariance Matrices in Reverberant Environments. *Appl. Sci.* **2022**, *12*, 4278. https://doi.org/10.3390/app12094278

Academic Editors: Yoshinobu Kajikawa, Cheng-Yuan Chang and Maciej Niedzwiecki

Received: 4 April 2022
Accepted: 20 April 2022
Published: 23 April 2022

1. Introduction

Direction-of-arrival (DOA) estimation has received more attention in the field of signal processing because it has a wide range of application. It can be applied to wireless communication, speech source separation, video conferencing and so on [1–6]. Over the years, many conventional methods have been successfully developed for the estimation task. The subspace-based methods are more popular among traditional methods for the outstanding performances. The multiple signal classification (MUSIC) [7] and the estimation of signal parameter via rotational invariance (ESPRIT) [8] are relatively eminent representative methods in subspace-based algorithms. The noise subspace is used to generate a spectrum in the MUSIC algorithm. The position where the spectral peak appears is the corresponding estimated DOA. Although it can guarantee relatively higher accuracy, the peak search is very time-consuming. Conversely, the ESPRIT method utilizes the signal sub-space for DOA estimation. The ESPRIT is time-saving at the expense of accuracy. I the beamforming-based approach, such as the minimum variance distortionless response (MVDR) [9], is another traditional method. However, the Rayleigh limit and low resolution are the shortcomings of this method. Generalized cross-correlation phase transform (GCC-PHAT) is also a commonly used traditional method, which requires the time difference of arrival (TDOA) between microphone pairs [10,11]. There is also a conventional called Maximum likelihood method [12]. It needs a multi-dimensional search, even though they can provide high estimation accuracy. Even worse, the global convergence can not be guaranteed. All these methods mentioned above can be used

in an anechoic condition. However, they generally suffer from problems such as high computational cost or performance degradation in noisy and reverberant environments.

In recent years, the development of deep learning has brought new inspiration to the DOA estimation. In order to resist reverberation and noise effectively, many deep neural networks (DNNs) have been proposed to estimate DOA. The most regular way to deal with the DOA task is to treat it as a regression task or a classification task. In terms of the regression, features extracted from the observation signals, which is related to location information, are directly mapped to the source position by deep learning [13–17]. For the classification, the position space will be divided into different regions. Each region corresponds to a class. Neural networks learn the relationship between input features and classes and then identify the class that a signal belongs to. In many studies and simulations, it is shown that when DOA is acted as a classification task, it can be estimated more accurately [18–21].

In the classification task, a vital step is to find proper input features for neural networks to learn the relationship between the features and labeled regions. Many different features have been proposed for DOA estimation, such as phases of signals [22], both phases and magnitudes of signals [23], real part and imaginary part of the signal [24], intensity vectors [25] and so on. Soumitro used signal phases as the inputs for a uniform linear array (ULA) [22]. However, elevations cannot be estimated simultaneously. It is not enough to locate the signal exactly with the ULA. Adavanne combined phase and amplitude spectra as inputs [23] and mapped them to the outputs. The input features are first mapped into space pseudo-spectra (SPS) and the DOAs in the two-dimensional polarized coordinates are estimated. Perotin utilized the intensity vectors of first-order ambisonic (FOA) signals as the inputs [25] to jointly estimate DOAs. However, the accuracy is relatively lower. As shown before, the spherical microphone array has not received much attention. However, it has many advantages for array signal processing. In particular, this array does not cause angular blurring due to the perfect symmetry of the sphere. Thus, this paper focused on the spherical microphone array, and three different input features and two different strategies are proposed for improving the accuracy of the localization task. The remainder of this paper is organized as follows. In Section 2, a data model of DOA estimation with spherical arrays is given. Our proposed method is described in Section 3. The architecture of the network is indicated in Section 4. Simulations and performance evaluation are shown in Section 5. Section 6 shows the computational complexity. Finally, conclusions are drawn in Section 7.

2. Data Model

A spherical microphone array with a radius r and L mutually independent and isotropic sensors in standard spherical coordinate [26] is considered. The center of the array coincides with the origin of the coordinate system and the microphones are located at the position $\boldsymbol{r}_l = r[\cos\varphi_l \sin\vartheta_l, \sin\varphi_l \cos\vartheta_l, \cos\vartheta_l]^T$, where ϑ_l and φ_l represent the elevation and azimuth of the l-th sensor, respectively, $l = 1, \cdots, L$. Figure 1 indicates the spherical coordinate system used for localization.

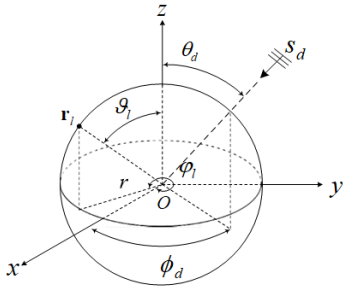


Figure 1. Spherical coordinate system.

An acoustic far-field signal $s_d(t)$ emitted from the location $\boldsymbol{r}_d = (r_d, \theta_d, \phi_d)$ is supposed. The distance between the source and the center of the spherical microphone array is r_d, and θ_d and ϕ_d mean the elevation and azimuth of the d-th source, respectively. Thus, the output vector $\mathbf{x}(t)$ of the spherical microphone array can be obtained as

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{v}(t), \tag{1}$$

where $\mathbf{x}(t) = [x_1(t), \cdots, x_L(t)]^T$, $\mathbf{A}$ is the steering matrix that represents the transfer functions from the source signals $\mathbf{s}(t) = [s_1(t), \cdots, s_D(t)]^T$ to the array microphones and holds the DOA information, $\mathbf{v}(t) = [v_1(t), \cdots, v_L(t)]^T$ is the noise vector, t is the time index, D represents the number of sources. When source signals are wideband ones, short-time Fourier transform (STFT) [27] must be considered to construct a signal model in the time–frequency domain from the spatial domain as follows

$$\mathbf{x}(\tau, k_h) = \mathbf{A}(k_h)\mathbf{s}(\tau, k_h) + \mathbf{v}(\tau, k_h), \tag{2}$$

where τ indicates the index of time frame, $k_h = 2\pi f_h / c$ is the wavenumber corresponding to the frequency bin f_h, h represents the frequency bin index, c is the speed of the sound, $\mathbf{x}(\tau, k_h) = [x_1(\tau, k_h), \cdots, x_L(\tau, k_h)]^T$, $\mathbf{v}(\tau, k_h) = [v_1(\tau, k_h), \cdots, v_L(\tau, k_h)]^T$, and $\mathbf{A}(k_h)$, the corresponding steering matrix [28], can be decomposed as

$$\mathbf{A}(k_h) = \mathbf{Y}(\Omega)\mathbf{B}(k_h r)\mathbf{Y}^H(\Phi), \tag{3}$$

where $\mathbf{Y}(\Omega) = [\mathbf{y}(\vartheta_1, \varphi_1), \cdots, \mathbf{y}(\vartheta_L, \varphi_L)]^T$ is the spherical harmonic matrix containing the information of the locations of all the microphones and $\mathbf{Y}(\Phi) = [\mathbf{y}(\theta_1, \phi_1), \cdots, \mathbf{y}(\theta_D, \phi_D)]^T$ is the spherical harmonic matrix only dependent on DOAs of D sources, $[\cdot]^T$ indicates transpose, $[\cdot]^H$ is the conjugate transpose, the spherical harmonics vector $\mathbf{y}(\theta, \phi)$ is expressed as

$$\mathbf{y}(\theta, \phi) = [Y_0^0(\theta, \phi), Y_1^{-1}(\theta, \phi), Y_1^0(\theta, \phi), Y_1^1(\theta, \phi), \cdots, Y_N^N(\theta, \phi)]^T, \tag{4}$$

where $Y_n^m(\theta, \phi)$ is the spherical harmonic function with order n and degree m described as

$$Y_n^m(\theta, \phi) = \sqrt{\frac{(2n+1)(n-m)!}{4\pi(n+m)!}} P_n^m(\cos\theta) e^{im\phi}, \tag{5}$$

where $i^2 = -1$, $P_n^m(x)$ is the associated Legendre function of order n and degree m. When m is positive, $P_n^m(x)$ is defined as

$$P_n^m(x) = (-1)^m (1 - x^2)^{\frac{m}{2}} \frac{d^m}{dx^m} P_n(x), \tag{6}$$

for the negative parts, it can be obtained from the positive parts as

$$P_n^{-m}(x) = (-1)^m \frac{(n-m)!}{(n+m)!} P_n^m(x), \tag{7}$$

where $P_n(x)$ is the associated Legendre polynomial given by

$$P_n(x) = \frac{1}{2^n n!} \frac{d^n}{dx^n} (x^2 - 1)^n. \tag{8}$$

N is the highest order of the spherical harmonic function, $n = 0, \cdots, N$ and $m = -n, \cdots, 0, \cdots, n$. $\mathbf{B}(k_h r) = diag\{b_0(k_h r), b_1(k_h r), b_1(k_h r), b_1(k_h r), \cdots, b_N(k_h r)\}$, where $b_n(k_h r)$ is the mode strength [29] defined as

$$b_n(k_h r) = \begin{cases} 4\pi i^n j_n(k_h r) & \text{open sphere} \\ 4\pi i^n j_n(k_h r) - \frac{j'_n(k_h r)}{o'_n(k_h r)} o_n(k_h r) & \text{rigid sphere} \end{cases}, \tag{9}$$

where $j_n(x)$ is spherical Bessel function of the first kind and $o_n(x)$ is spherical Hankel function of the second kind. $j'_n(x)$ and $o'_n(x)$ are their derivatives, respectively

The model in (2) can be rewritten as

$$\mathbf{x}(\tau, k_h) = \mathbf{Y}(\Omega)\mathbf{B}(k_h r)\mathbf{Y}^H(\Phi)\mathbf{s}(\tau, k_h) + \mathbf{v}(\tau, k_h). \tag{10}$$

The model can be transformed into the spherical harmonic domain by left-multiplying $\mathbf{Y}^H(\Omega)$. Due to the orthogonality principle of spherical harmonics for uniform or nearly uniform sampling, it can be known that $\mathbf{Y}^H(\Omega)\mathbf{Y}(\Omega) = \mathbf{I}_{(N+1)^2}$ [30], where $\mathbf{I}_{(N+1)^2}$ is an $(N+1)^2 \times (N+1)^2$ identity matrix. The new model in the spherical harmonic domain is

$$\mathbf{x}_{nm}(\tau, k_h) = \mathbf{B}(k_h r)\mathbf{Y}^H(\Phi)\mathbf{s}(\tau, k_h) + \mathbf{v}_{nm}(\tau, k_h), \tag{11}$$

where $\mathbf{x}_{nm}(\tau, k_h) = \mathbf{Y}^H(\Omega)\mathbf{x}(\tau, k_h)$ and $\mathbf{v}_{nm}(\tau, k_h) = \mathbf{Y}^H(\Omega)\mathbf{v}(\tau, k_h)$. Further, (11) can be left multiplied with $\mathbf{B}^{-1}(k_h r)$.

$$\hat{\mathbf{x}}_{nm}(\tau, k_h) = \mathbf{Y}^H(\Phi)\mathbf{s}(\tau, k_h) + \hat{\mathbf{v}}_{nm}(\tau, k_h), \tag{12}$$

where $\hat{\mathbf{x}}_{nm}(\tau, k_h) = \mathbf{B}^{-1}(k_h r)\mathbf{x}_{nm}(\tau, k_h)$, $\hat{\mathbf{v}}_{nm}(\tau, k_h) = \mathbf{B}^{-1}(k_h r)\mathbf{v}_{nm}(\tau, k_h)$. In this model, the steering matrix $\mathbf{Y}^H(\Phi)$ is irrelevant to the information of the microphones and mode strength.

3. The Proposed Method

This section introduces the input features of the proposed method for the DOA task. As shown before, three different covariance matrices are considered as the input features for the neural network. The first covariance matrix is the covariance matrix of the received signal of the microphone array in the time–frequency domain (TFD). The second one is the covariance matrix of the signal in the spherical harmonic domain (SHD). The third one is the covariance matrix of the signal in the azimuth–elevation domain (AED). All the covariance matrices are to be introduced in detail.

3.1. TFD Matrix

The received signal of the microphone array in the time–frequency domain is the output vector $\mathbf{x}(\tau, k_h)$ of the microphone array. The covariance matrix of the $\mathbf{x}(\tau, k_h)$ can be obtained as

$$\mathbf{M}_1 = \frac{1}{N_F}\frac{1}{N_T}\sum_{h=1}^{N_F}\sum_{\tau=1}^{N_T}\mathbf{x}(\tau, k_h)\mathbf{x}^H(\tau, k_h), \tag{13}$$

where N_F is the number of the frequency bins, and N_T time frames are considered. In practical applications, the covariance matrix can also be calculated in this way. Under each frequency bin, the covariance matrix of the $\mathbf{x}(\tau, k_h)$ can be obtained along the time index. The final step is to average all the matrices with the total frequency bins.

3.2. SHD Matrix

It can be known that the data model in (12) has filtered out the effects of the microphone array and mode strength to some extent. Thus, this model can contain the angle information more efficiently. Thus, the covariance matrix of the model in the spherical harmonic domain can be got by the same way as the TFD matrix

$$\mathbf{M}_2 = \frac{1}{N_F}\frac{1}{N_T}\sum_{h=1}^{N_F}\sum_{\tau=1}^{N_T}\hat{\mathbf{x}}_{nm}(\tau,k_h)\hat{\mathbf{x}}_{nm}{}^H(\tau,k_h). \tag{14}$$

3.3. AED Matrix

A mapping matrix between the spherical harmonic function and Fourier series was found in our previous work [31]. Its dimension is $(2N+1)^2 \times (2N+1)^2$. The mapping matrix has the function to transform the steering vector in (12) into the Kronecker product of two Fourier series vectors. Each vector has a Vandermonde structure only dependent on azimuths or elevations, respectively,

$$\mathbf{y}(\theta,\phi) = \mathbf{E}[\mathbf{f}(\phi)\otimes\mathbf{f}(\theta)], \tag{15}$$

where operator $\otimes$ represents the Kronecker product. The $\mathbf{f}(\phi)$ and $\mathbf{f}(\theta)$ can be expressed as

$$\mathbf{f}(\phi) = [e^{iN\phi},\cdots,1,\cdots,e^{-iN\phi}]^T, \tag{16}$$

$$\mathbf{f}(\theta) = [e^{-iN\theta},\cdots,1,\cdots,e^{iN\theta}]^T. \tag{17}$$

Moreover, the values in matrix $\mathbf{E}$ do not change with the positions of sound sources. Thus, (15) can be left multiplied with the generalized inversion of matrix $\mathbf{E}$, and it can obtain

$$\mathbf{g}(\theta,\phi) = \mathbf{E}^{\dagger}\mathbf{y}(\theta,\phi) = \mathbf{f}(\phi)\otimes\mathbf{f}(\theta). \tag{18}$$

According to (12) and (18), the novel model can be obtained from (12) by left multiplying $(\overline{\mathbf{E}})^{\dagger}$

$$\mathbf{x}_{\theta\phi}(\tau,k_h) = \overline{\mathbf{G}}(\Phi)\mathbf{s}(\tau,k_h) + \mathbf{v}_{\theta\phi}(\tau,k_h), \tag{19}$$

where $\mathbf{x}_{\theta\phi}(\tau,k_h) = (\overline{\mathbf{E}})^{\dagger}\hat{\mathbf{x}}_{nm}(\tau,k_h)$ and $\mathbf{v}_{\theta\phi}(\tau,k_h) = (\overline{\mathbf{E}})^{\dagger}\hat{\mathbf{v}}_{nm}(\tau,k_h)$. $\overline{[.]}$ is the operation that conjugates all the elements in the matrix. $\mathbf{G}(\Phi) = [\mathbf{g}(\theta_1,\phi_1),\cdots,\mathbf{g}(\theta_D,\phi_D)]$. Compared with $\hat{\mathbf{x}}_{nm}(\tau,k_h)$, $\mathbf{x}_{\theta\phi}(\tau,k_h)$, which is determined by the azimuth and elevation to the greatest extent, further efficiently removes information irrelevant to location information; therefore, it can be seen that the signal is transformed from the time–frequency domain to the azimuth–elevation domain. Thus, one can obtain the covariance matrix of the $\mathbf{x}_{\theta\phi}(\tau,k_h)$ with the same method

$$\mathbf{M}_3 = \frac{1}{N_F}\frac{1}{N_T}\sum_{h=1}^{N_F}\sum_{\tau=1}^{N_T}\mathbf{x}_{\theta\phi}(\tau,k_h)\mathbf{x}_{\theta\phi}{}^H(\tau,k_h), \tag{20}$$

The last step is to assemble the real parts and the imaginary parts to obtain the final input features. Table 1 indicates the dimension of the three different covariance matrices.

Table 1. The dimensions of the three different covariance matrices.

	TFD	SHD	AED
Dimension	$2L \times L$	$2(N+1)^2 \times (N+1)^2$	$2(2N+1)^2 \times (2N+1)^2$

Figure 2 is the flowing chart of acquiring the input features. From the figure and theoretical analysis, it can be found that the AED best carries the information of the location, which will be the most efficient input features at a theoretical level. Algorithm 1 shows the steps of the algorithm for palpability.

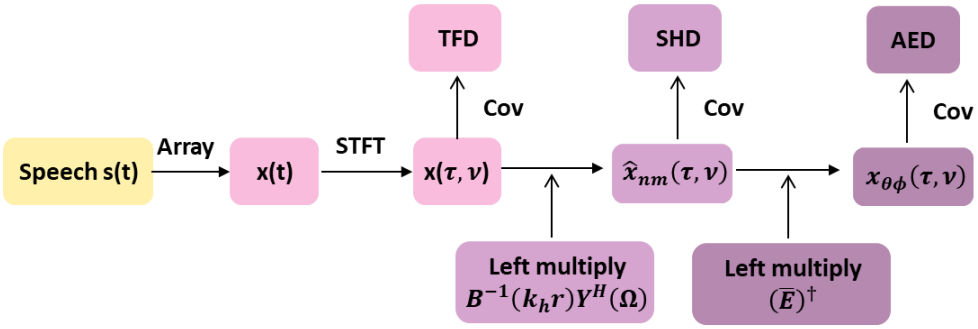


Figure 2. The flowing chart of obtaining the input features. Cov represents the operation to acquire the covariance matrix of the signal.

Algorithm 1 The algorithm of proposed input features

Input: received signal $\mathbf{x}(t)$
Output: TFD, SHD, AED

Get $\mathbf{x}(\tau, k_h)$ by using the STFT of the received signal $\mathbf{x}(t)$.
$\mathbf{x}_{nm}(\tau, k_h) = \mathbf{Y}^H(\Omega)\mathbf{x}(\tau, k_h)$.
$\hat{\mathbf{x}}_{nm}(\tau, k_h) = \mathbf{B}^{-1}(k_h r)\mathbf{x}_{nm}(\tau, k_h)$.
$\mathbf{x}_{\theta\phi}(\tau, k_h) = (\bar{\mathbf{E}})^{\dagger}\hat{\mathbf{x}}_{nm}(\tau, k_h)$
$\mathbf{M}_1 = \text{zeros}(L, L)$, $\mathbf{M}_2 = \text{zeros}((N+1)^2, (N+1)^2)$, $\mathbf{M}_3 = \text{zeros}((2N+1)^2, (2N+1)^2)$.
for $k_h = 1 : N_F$
for $\tau = 1 : N_T$
$\mathbf{M}_1 = \mathbf{M}_1 + \text{cov}(\mathbf{x}(\tau, k_h))$
$\mathbf{M}_2 = \mathbf{M}_2 + \text{cov}(\hat{\mathbf{x}}_{nm}(\tau, k_h))$
$\mathbf{M}_3 = \mathbf{M}_3 + \text{cov}(\mathbf{x}_{\theta\phi}(\tau, k_h))$
end for
end for
$\mathbf{M}_1 = \frac{1}{N_F N_T}\mathbf{M}_1$
$\mathbf{M}_2 = \frac{1}{N_F N_T}\mathbf{M}_2$
$\mathbf{M}_3 = \frac{1}{N_F N_T}\mathbf{M}_3$
TFD $\leftarrow [\text{real}(\mathbf{M}_1); \text{image}(\mathbf{M}_1)]$.
SHD $\leftarrow [\text{real}(\mathbf{M}_2); \text{image}(\mathbf{M}_2)]$.
AED $\leftarrow [\text{real}(\mathbf{M}_3); \text{image}(\mathbf{M}_3)]$.

4. Learning Strategies and Network Architectures

Two different learning strategies and their corresponding network structures are introduced in this section. As shown before, the DOA estimation is regarded as a classification task. The task can be finished with two strategies.

4.1. Regular Strategy

The estimation is treated as a classification task. The whole space is divided into different classes and the emitted signal detects which class the signal belongs to. The elevation space, ranging from 30° to 150°, is divided into 13 different classes with the interval of 10°. Similarly, the azimuth space, ranging from 0° to 360°, is divided into 36 classes with the interval of 10°. In total, 468 different classes can be obtained. Moreover, each class is granted a label to distinguish it. For each class, a large number of samples are generated for the network to learn and give the network the ability to localize the sound source. For the classification, a convolutional neural network is utilized for the task. The network is composed of three convolutional layers and two fully connected layers. For the convolutional layers, the size of the convolution cores in these three convolutional layers is 2×2. Moreover, the number of filters in all layers is fixed at 8. All the convolutional

layers are activated by the rectified liner unit (ReLU) activation [32]. A flattened layer follows the last convolutional layer to reshape the output of the convolutional layers. The following components are two fully connected (FC) layers. The first layer with N_H nodes is activated by the ReLU activation. The second layer with N_C nodes uses softmax for the activation. N_C is the number of classes for the task. The softmax function [33] helps to obtain the posterior probabilities of all the candidates. Finally, according to the principle of maximum posterior probability, the position where the maximum probability occurs is the estimated DOA. Figure 3 demonstrates the details of the network for the estimation task. During learning, the cross-entropy [34] is chosen as the loss function because its derivation is simpler and it is only related to the probability of the correct class. Furthermore, the loss is able to conveniently derive the input of the softmax activation layer. Adam [35] is adopted for the stochastic optimization. The details of the CNN are shown in Figure 3.

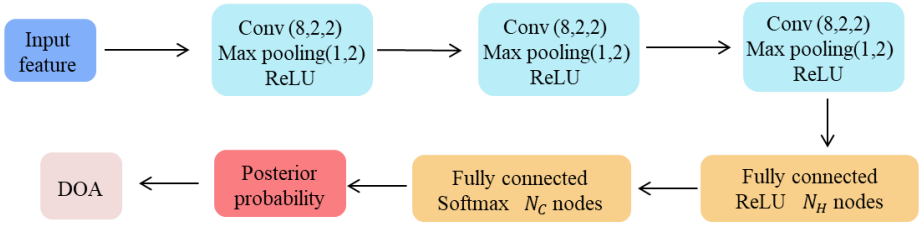


Figure 3. The process for DOAs estimation and architecture in detail. Conv (x, y, z) indicates that the size of the convolution core is $y \times z$, and the number of the filters is x. N_H is the number of the nodes in the first FC layers. N_C is the number of the nodes in the second FC layer, which represents the number of classes.

4.2. Segmentation Strategy

As shown in Section 4.1, there are 468 different classes in the regular strategy. Too many classes can bring interference to learning and testing. Thus, this part proposes the segmentation strategy for predigestion. For the azimuth, the entire space is considered, which is a major contributor to too many classes. The azimuth space can be processed with the segmentation strategy. The entire space is equally divided into Q sub-spaces. Thus, the task is finished in parallel with Q independent networks. This strategy can decompose the excessive number of classes in the task into Q parts so that the number of the total classes in each part is reduced, thereby reducing the interference induced by classes. Figure 4 shows the example of the space division when $Q = 3$, and the experiments in this paper are conducted in the case of $Q = 3$. The color-rendered part is the corresponding sub-space, and the white area is the irrelevant space compared to the coloring space.

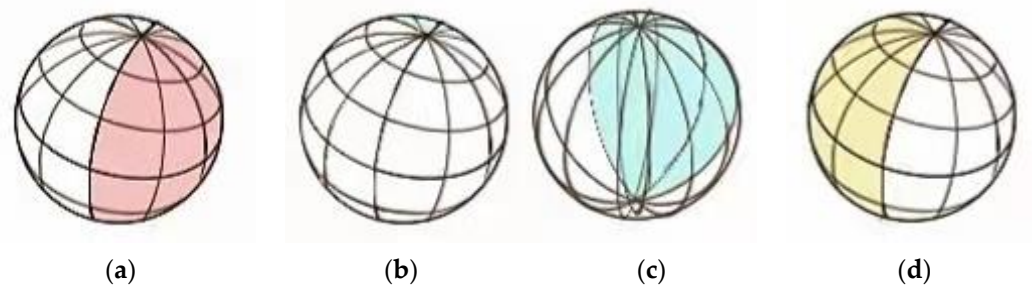


Figure 4. Schematic diagram of space division. (**a**) is the first sub-space, which ranges from 0° to 120°. (**b**) is the second sub-space, which ranges from 120° to 240°. (**c**) is a perspective view of (**b**). (**d**) is the third sub-space, which ranges from 240° to 360°.

Take Figure 4a as an example. The pink part is the first sub-space used for the first CNN. The space range for the pink part is from 0° to 120°, thus, 169 classes can be obtained for the first CNN. As the other blank parts are irrelevant spaces, a new class is needed in this subtask. This class represents irrelevant sound, which means that the signal comes

from irrelevant areas. This feature should be used as the input of the other two networks to determine the location, the signal does not appear in this sub-space. Thus, 170 classes are considered for the first sub-CNN. Similarly, the second sub-space, ranging from 120° to 240°, also has 170 classes. The third sub-space also has this, whose range is from 240° to 360°.

Figure 5 shows the architecture of the network in the segmentation strategy with $Q = 3$. The specific structure of the sub-CNN is the same as the network structure of the CNN in the regular strategy to better compare the effectiveness of strategies.

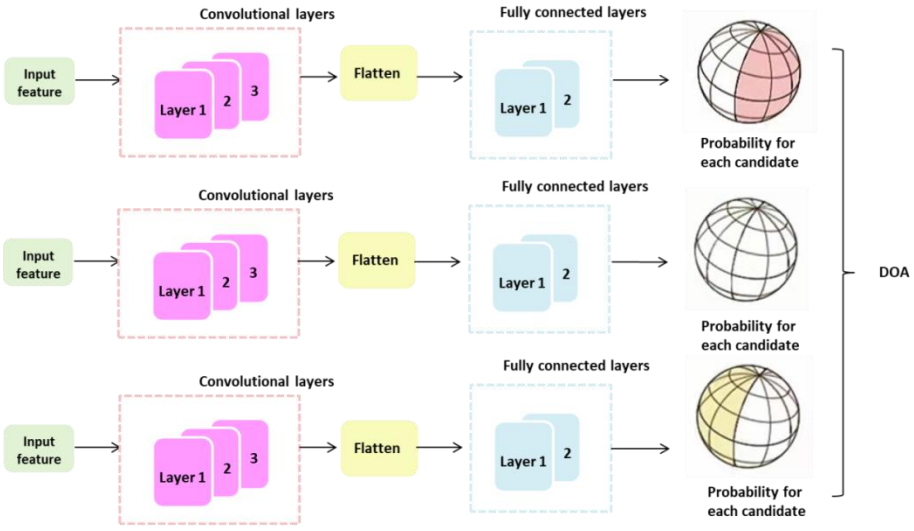


Figure 5. The process for DOA estimation task and architecture in detail for the segmentation strategy.

For testing, the input feature of one signal is sent to three sub-networks for detection at the same time, and the three networks output all the posterior probabilities of the whole space and judge the position of the signal through the principal of the maximum posterior probability. In particular, when the maximum probability appears in the irrelevant class, the entire corresponding sub-space will no longer be considered.

5. Simulations and Evaluation

This section introduces the setups for the simulations and the evaluation for the performance. The audio settings followed are similar to those in [28]. A Hanning window is used with the length 256 and an overlap of 75% between adjacent frames. The sample frequency is 16 kHz. Moreover, 256-point FFT is considered. The frequency bins ranging from 500 to 3875 Hz are taken into consideration to have sufficient mode strengths and avoid aliasing, which results in the number of the frequency bins N_F being 55. All these parameters concern the audio settings. The signal can be processed more effectively with the help of these parameters. Furthermore, the appropriate frequency range can be selected in order to have sufficient mode strengths and avoid aliasing with these parameters. Moreover, the highest order of spherical harmonics is 4. A rigid spherical microphone array with $L = 32$ and $r = 4.2$ cm is considered. The highest order of spherical harmonics and the number of the sensors directly influence the dimension of the input features. Thus, the dimension of the TFD matrix is 64×32. The dimension of the SHD matrix is 50×25. The dimension of the AED matrix is 162×81.

5.1. Dataset

In this section, the simulated and real datasets for the algorithm are introduced. For training, a large number of data are needed; however, in practice, it is time-consuming and less cost-effective to sample the data. Thus, simulating data can be taken into consideration for comprehensive training. A variety of simulated environments can be obtained by using

SMA response (SMIR) [36]. The SMIR is decided by the size of the room, the locations of array microphones and sound sources and the reverberation time (RT_{60}). Figure 6 shows the flow chart of the process of the simulated dataset generation.

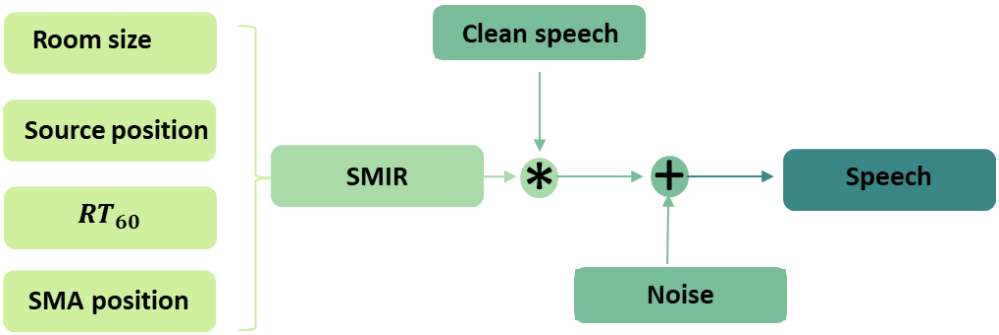


Figure 6. The process for obtaining the speech with position information and reverberation.

In terms of the room size, six rooms are considered. The six rooms have different sizes except for a fixed height of 3 m. Four rooms are used for training, and all the rooms are utilized for testing. RT_{60} can be chosen between 0.2 and 1 s. Eigenmike [37], a rigid SMA of radius of 4.2 cm with $L = 32$, $N = 4$, is used as the array for generating the signals. The distance between each source and the center of the SMA is at least 1 m. Moreover, the source and SMA positions are at a minimum distance of 50 cm away from all the walls in the room. Clean speeches are chosen for generating the dataset. Then, all the SMIRs and the source signals are convolved to obtain new signals. In terms of noises, white Gaussian noise is considered. Moreover, voice activity detection (VAD) [38] is performed on the clean speeches to filter out the silent frames and only maintain the speech frames. The method utilized for generating class labels is the same as that in [39] to treat DOA estimation as a classification task.

5.1.1. Training Dataset

In terms of the room size, four rooms are considered. The size of the room is measured in meters. The four rooms have different sizes except for a fixed height of 3 m. The room sizes are: Room 1 (3 m × 5 m × 3 m); Room 2 (6 m × 8 m × 3 m); Room 3 (7 m × 10 m × 3 m); Room 4 (8 m × 9 m × 3 m). As shown before, the elevation space and azimuth space are both divided with the interval of 10°. Each class is granted a label. For each DOA class, 125 SMIRs can be obtained by combining different parameters. Eight different speech signals are randomly selected from the Librispeech dev corpus [40] to convolve all the SMIRs. Thus, from the four rooms with different sizes, a total of 4000 samples for each DOA class can be acquired. The signal to noise ratio (SNR) is randomly chosen between 0 and 30 dB. A validation dataset is also needed, whose size is 10% of that of the training dataset.

5.1.2. Testing Dataset

The test dataset is divided into two different parts. The first part is the simulated test dataset, which is similar to the training dataset but differs in terms of SNR, RT_{60} and the source localizations. In total, 265 different DOA positions are selected randomly. In addition to the four rooms already mentioned in the training set, there are two additional rooms: Room 5 (4 m × 6 m × 3 m) and Room 6 (9 m × 7 m × 3 m). The rule of selecting the position of the SMA is the same as that for the training dataset. For better analysis of the relationship between the results and the parameters, RT_{60} and SNR are no longer selected randomly. RT_{60} varies from 0.2 s to 1 s in a step size of 0.2 s. The SNR changes from 0 dB to 30 dB with a step of 5 dB.. Similarly, two speech recordings are selected for the convolution with SMIR. However, this time, the recordings are selected from the Librispeech test corpus [40], not the Librispeech dev corpus. The second is the real test dataset. The LOCATA dataset [41] is

used as a real dataset. This is the IEEE-AASP challenge on sound source localization and tracking. From this, task 1 is chosen to localize the direction of the single static loudspeaker. The recordings are conducted in a room of (7.1 m × 9.8 m × 3 m) with $RT_{60} = 0.55$ s. Table 2 indicates the parameters for the simulated datasets.

Table 2. Parameters for the simulated datasets.

Parameter	Training Stage	Testing Stage
DOA	θ: step-10°; ϕ: step-10°	Randomly
SNR	Randomly	step-5dB
RT_{60}	Randomly	step-0.2s
Speech	Librispeech Dev	Librispeech Test
Noise	White Gaussian Noise	White Gaussian Noise
Room	Room 1–4	Room 1–6

The parameters mentioned above are very important for generating the dataset. Different room sizes are used to enrich the information on scene and are beneficial for the algorithm to apply to more rooms. The white Gaussian noise is used as the interference to verify the robustness of the algorithm to interference. SNR and RT_{60} are utilized to enrich the complexity of the environments. Moreover, during the testing stage, these two parameters can help to show the relationship between the results and the conditions, which is effective for verifying the feasibility of the algorithm.

5.2. Methods for Comparison

Spherical harmonic MUSIC (SH-MUSIC) method [42], direct-path domain test method for MUSIC algorithm (DPD-MUSIC) [28,43] and FOA [44] method are used for the performance comparison.

5.2.1. SH-MUSIC

The SH-MUSIC algorithm is the MUSIC method conducted in the spherical harmonic domain. A spectrum is computed for all the possible candidates of DOAs, and the peak is found to give the corresponding DOA of the source

$$(\hat{\theta}_d, \hat{\phi}_d) = \arg \max_{(\theta,\phi)\in\mathfrak{L}} \frac{1}{\mathbf{y}^H(\theta,\phi)\mathbf{U}_n\mathbf{U}_n^H\mathbf{y}(\theta,\phi)}, \tag{21}$$

where $\mathfrak{L}$ is the DOA search grid set containing all the possible DOA candidates, $\mathbf{U}_n$ is the noise sub-space decomposed from the covariance matrix of the array signals using eigenvalue decomposition [45]. The covariance matrix can be computed from (14).

5.2.2. DPD-MUSIC

In order to effectively resist reverberation, a direct-path dominance (DPD) test is often applied to select those time–frequency (TF) bins that contain more direct components. The test can be expressed as

$$\mathfrak{T}_{DPD} = \{(\tau, k_h) : \frac{\sigma_1(\tau,k_h)}{\sigma_2(\tau,k_h)} \geqslant \lambda_{DPD}\}, \tag{22}$$

where λ_{DPD} is the threshold for the DPD test to choose the TF-bins, $\sigma_1(\tau,k_h)$ and $\sigma_2(\tau,k_h)$ are the largest and second-largest singular values of the covariance matrix $\tilde{\mathbf{R}}_{nm}(\tau,k_h)$ obtained using a rectangular window over the time and frequency indices.

$$\tilde{\mathbf{R}}_{nm}(\tau,k_h) = \frac{1}{J_\tau J_v}\sum_{\iota_\tau=0}^{J_\tau-1}\sum_{\iota_v=0}^{J_v-1}\hat{\mathbf{x}}_{nm}(\tau+\iota_\tau,k_h+\iota_v)\hat{\mathbf{x}}_{nm}^H(\tau+\iota_\tau,k_h+\iota_v) \tag{23}$$

The window contains J_τ time frames and J_v frequency bins. Based on the selected TF bins, SH-MUSIC method is still used for the sound localization task [28,43].

5.2.3. FOA

Ref. [44] shows that the first four channels of the FOA are considered. This step involves convolving the signals of the last three channels with the one in the first channel, assembling the convolved result, and then taking the real and imaginary parts of the new result as input. In order to better compare the effectiveness of different inputs, all methods are performed by using the same network structure.

5.3. Measure

In order to evaluate the performance, the gross error (GE) is used for the measure stage. The GE is the percentage of estimation that does not fall into an allowed threshold from the original value. Thus, for the simulated dataset, the GE is used to evaluate the performance. As for the LOCATA dataset, mean deviation and standard deviation [41] are utilized for the evaluation. The GE is defined as

$$\mathrm{GE}_\theta = \frac{1}{N_K}\sum_{\kappa=1}^{N_K}[I_e(|\theta_\kappa - \hat{\theta}_\kappa| - \lambda)], \tag{24}$$

$$\mathrm{GE}_\phi = \frac{1}{N_K}\sum_{\kappa=1}^{N_\kappa}[I_e(|\phi_\kappa - \hat{\phi}_\kappa| - \lambda)], \tag{25}$$

$$\mathrm{GE} = \frac{1}{N_K}\sum_{\kappa=1}^{N_K}[I_e(\Delta((\theta_\kappa, \phi_\kappa), (\hat{\theta}_\kappa, \hat{\phi}_\kappa)) - \lambda)], \tag{26}$$

where N_K is the number of DOAs estimated, $(\theta_\kappa, \phi_\kappa)$ and $(\hat{\theta}_\kappa, \hat{\phi}_\kappa)$ are the actual and the estimated DOA of the κth source, λ is the threshold that represents the maximum acceptable error, Δ is the angular distance and it is defined as

$$\Delta[(\theta, \phi), (\hat{\theta}, \hat{\phi})] = \arccos[\sin\hat{\theta}\sin\theta + \cos\hat{\theta}\cos\theta\cos(\hat{\phi} - \phi)], \tag{27}$$

$I_e(x)$ is

$$I_e(x) = \begin{cases} 1 & \text{for } x \geqslant 0 \\ 0 & \text{for } x < 0 \end{cases}. \tag{28}$$

As introduced before, for the simulated dataset, the GE is used for the evaluation. $\lambda = 10, 15, 20$ is considered. Table 3 demonstrates the GE for the azimuth estimation with different thresholds. Table 4 illustrates the GE for the elevation estimation with different thresholds. The symbol '-S' in the tables means that this task is finished with the segmentation strategy. The GE value shown in the two tables is an average over all the combinations of SNR and RT_{60}.

Table 3. GE for the azimuth estimation.

Method	λ = 20	λ = 15	λ = 10
MUSIC	52.68%	65.02%	74.56%
DPD-MUSIC	38.77%	41.77%	44.34%
FOA	16.77%	23.96%	28.01%
TFD	21.56%	26.67%	30.45%
TFD-S	18.23%	24.63%	27.89%
SHD	16.98%	23.63%	28.77%
SHD-S	14.84%	19.62%	25.33%
AED	**13.05%**	**18.08%**	**25.08%**
AED-S	**9.96%**	**14.77%**	**21.86%**

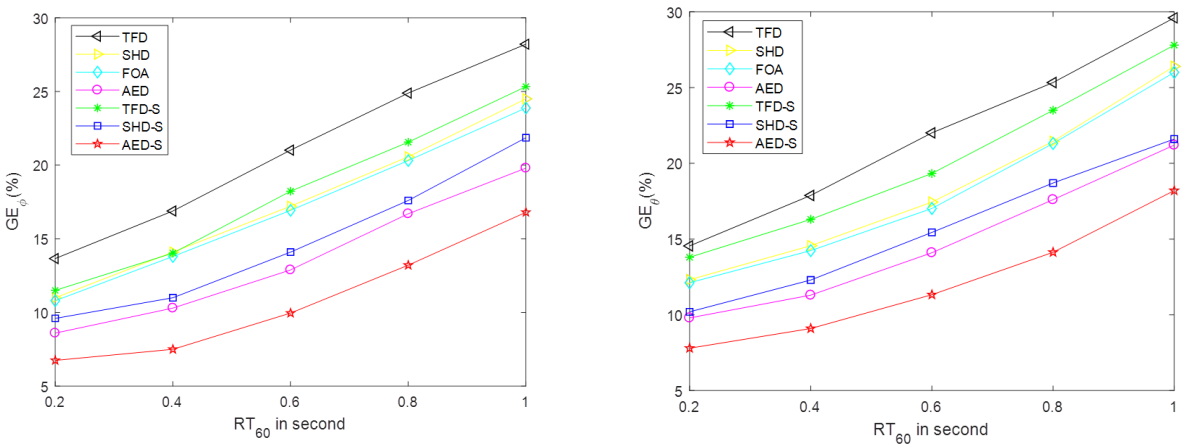
Table 4. GE for the elevation estimation.

Method	λ = 20	λ = 15	λ = 10
MUSIC	51.56%	63.46%	72.41%
DPD-MUSIC	36.41%	38.55%	45.06%
FOA	18.24%	24.01%	28.96%
TFD	22.01%	27.78%	31.30%
TFD-S	19.33%	26.45%	29.41%
SHD	18.02%	23.88%	28.46%
SHD-S	16.63%	20.31%	27.13%
AED	**13.43%**	**19.27%**	**26.11%**
AED-S	**11.33%**	**17.31%**	**25.05%**

From Table 3, it can be seen that the proposed input feature, the AED matrix, is the most efficient input feature for the azimuth estimation task. Furthermore, when the input features are fixed, compared to the regular strategy, the segmentation strategy has higher accuracy in the test of the azimuth estimation. Thus, the combination of the AED matrix and the segmentation strategy is the best choice for the azimuth.

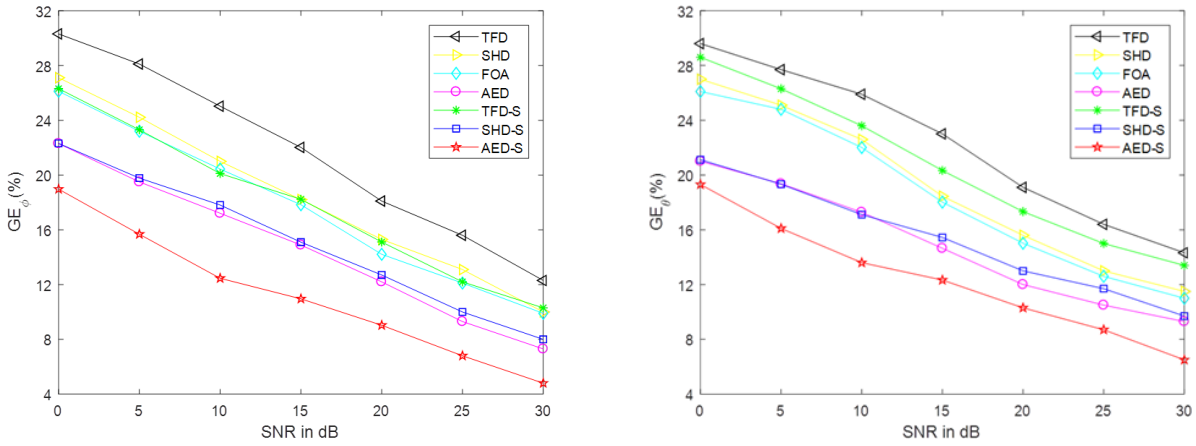
It can be known from Table 4 that the proposed input feature, the AED matrix, is also the most efficient input feature for the elevation estimation task. Moreover, when the input features are fixed, compared to the regular strategy, the segmentation strategy still has higher accuracy in the test of the elevation estimation. Although the improvement is not as remarkable as that in the azimuth estimation task, the segmentation strategy does bring improvement. Therefore, the proposed method can effectively improve the accuracy for both the azimuth estimation and elevation estimation.

The relationships between the results and the parameters are shown in Figures 7–9. From Table 3, Table 4, it can be known that the SH-MUSIC method and the DPD-MUSIC algorithm perform worse than the deep learning methods. Thus, these two methods are removed from the experiments for exploring the relationship between results and influencing factors. The relationship between the results and RT_{60} is given in Figure 7 with the threshold $\lambda = 20$. Figure 7a shows the GE for azimuth against RT_{60}, Figure 7b shows the GE for elevation against RT_{60}. Figure 8 shows the relationship between the results and SNR with the threshold $\lambda = 20$. Figure 8a shows the GE for azimuth against SNR, and Figure 8b shows the GE for elevation against SNR. Figure 9 indicates the relationship between the results and parameters, which is given in with the threshold $\lambda = 10$. Figure 9a shows the GE against RT_{60}, and Figure 9b shows the GE against SNR.



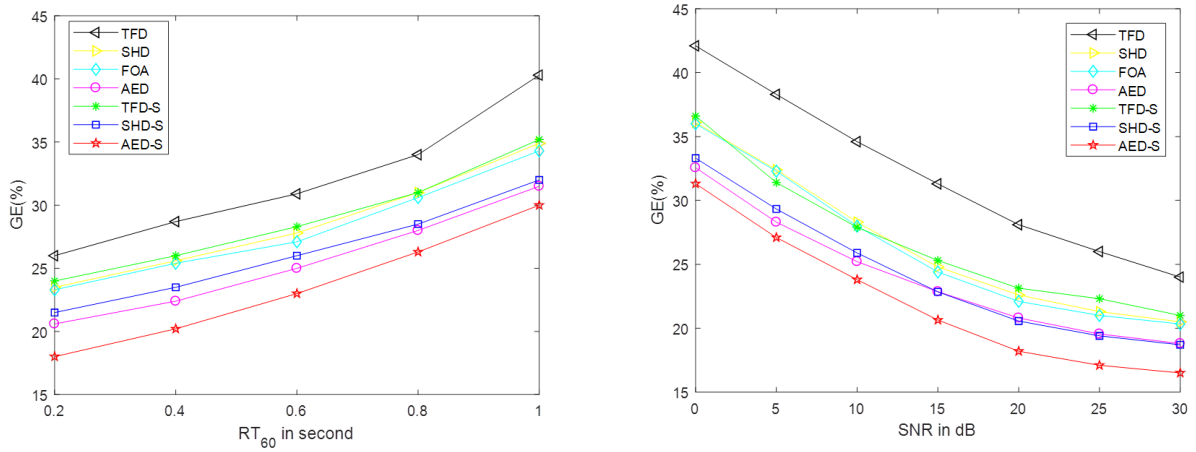
(**a**) (**b**)

Figure 7. GE for estimation against RT_{60} with the threshold $\lambda = 20$. (**a**) is the GE for azimuth against RT_{60}. (**b**) is the GE for elevation against RT_{60}.



(**a**) (**b**)

Figure 8. GE for estimation against SNR with the threshold $\lambda = 20$. (**a**) is the GE for azimuth against SNR. (**b**) is the GE for elevation against SNR.



(**a**) (**b**)

Figure 9. GE for estimation with the threshold $\lambda = 10$. (**a**) is the GE against RT_{60}. (**b**) is the GE against SNR.

From Figure 7, it can be obtained that the proposed input feature, the AED matrix, is the best input feature for improving accuracy. Furthermore, the segmentation strategy is also effective. Thus, the combination of the AED matrix and the segmentation strategy can help to resist reverberation effectively.

From Figure 8, it can be found that the proposed input feature, the AED matrix, is also the best input feature for reducing the error rate. Furthermore, the segmentation strategy is still effective. Thus, the combination of the AED matrix and the segmentation strategy can help to effectively resist noise.

From Figure 9, it can be seen that the proposed input feature still performs best even with the lower threshold. Thus, it can be concluded that the proposed method can improve accuracy in all the simulated conditions.

5.4. Results in the Real Dataset

The first task in the LOCATA challenge is used for real data test. It deals with the localization of a single static loudspeaker using a static SMA. The mean and standard deviation represent the performance errors showed in Table 5. From the table, it can be seen that the proposed method has lower mean deviation and lower standard deviation for the real dataset test.

Table 5. Mean and standard deviation for the LOCATA TASK-1.

Method	Mean Deviation	Standard Deviation
FOA	10.55°	10.19°
TFD	11.02°	10.88°
TFD-S	10.77°	10.54°
SHD	10.36°	9.89°
SHD-S	10.14°	8.83°
AED	9.81°	8.38°
AED-S	9.46°	7.98°

6. Computational Complexity

This section shows the comparison of the computational complexity and elapsed time. As shown before, a spherical microphone array is considered. Thus, the spherical harmonic order needs to be taken into consideration. Thus, only DPD-MUSIC and the proposed method are considered.

6.1. DPD-MUSIC

Ref. [46] shows the run-time complexity of the narrowband signal estimation using SH-MUSIC method in the spatial domain. The complexity is $O((N+1)^6 + (N+1)^2 N_c)$, where N_C is the total number of the DOA search grids. In terms of DPD-MUSIC, each rectangular window of time–frequency bins, whose size is $J_\tau \times J_{k_h}$, needs the eigenvalue decomposition. The number of the windows is $(N_T - J_\tau + 1) \times (N_F - J_{k_h} + 1)$. The J_τ and J_{k_h} are typically small compared to N_T and N_F [21,28]. Thus, the amount of the rectangular windows is $O(N_T N_F)$. Thus, the complexity of DPD-MUSIC is defined as $O(N_T N_F (N+1)^6 + (N+1)^2 N_\theta N_\phi)$.

6.2. Proposed Method

For the proposed method, it can be known that for a traditional CNN, about 90% of the parameters of the total network are in the FC layers [47]. Hence, the computational complexity can be decided by the FC layers while ignoring the influence of the convolutional layers. It can be understood that the input feature is directly given to the first FC layer. Thus, the first part of the computational complexity is decided by the size of the input feature and the number of the nodes in the first FC layer. Similarly, the second part is decided by the output of the first FC layer and the number of the nodes in the second FC layer. Thus, the complexity of the proposed method is $O(N_H N_I + N_H N_C)$, and N_I indicates the dimension of the input feature. Moreover, the proposed method has three different input features. From Table 1, the dimension can be obtained. Thus, the complexities are $O(N_H(2M^2) + N_H N_C)$ for TFD, $O(N_H(2(N+1)^4) + N_H N_C)$ for SHD and $O(N_H(2(2N+1)^4) + N_H N_C)$ for AED. From the theorical analysis, it can be seen that the proposed method has lower computational complexity than the DPD-MUSIC.

6.3. Elapsed Time Comparison

The elapsed time comparison among the algorithms for the test stage is shown in this section. The elapsed time is calculated from acquiring the feature to obtaining the DOA. For a fair comparison, all the experiments are conducted in the same system with Core (TM) i5-4800S processor manufactured by Inter (R), RAM of 16 GB, 64-bit instruction set. A clock speed of 2.81 GHz is used for execution. Figure 10 shows the elapsed time comparison in the form of a bar chart. It can be observed that the computational complexity is reduced. From the results of the bar chart and theoretical analysis, it can be known that the proposed method reduces the computational complexity

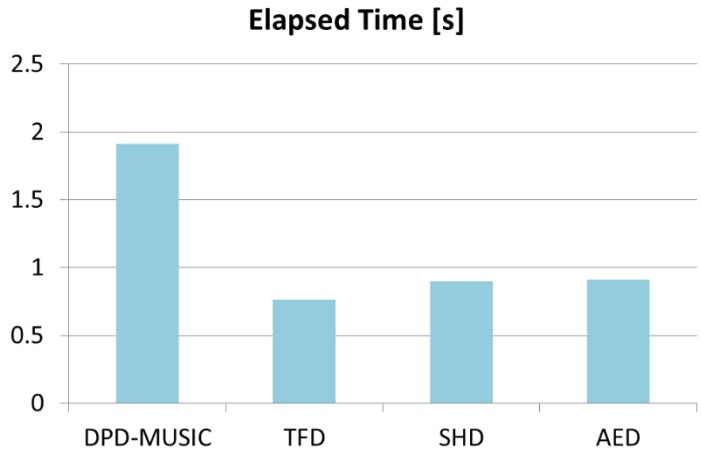


Figure 10. Elapsed time for different methods.

7. Conclusions

This paper proposed three different covariance matrices as the input features and two different learning strategies for the DOA task. For the matrices, there is a progressive relationship among the three covariance matrices. The second matrix can be obtained by processing the first matrix, and the third matrix can be obtained by processing the second matrix. Compared with TFD, SHD effectively filters out the effects of the microphone array and mode strength to some extent. Compared with SHD, AED more efficiently removes information irrelevant to location information. In terms of the strategies, the first strategy is a regular learning strategy while the second strategy is to split the task into three parts to be performed in parallel. Experiments were conducted both on the simulated and real dataset to show that the combination of the AED features and the segmentation strategy can achieve the best performance. Moreover, through theoretical analysis and elapsed time, it can be observed that the proposed method has lower computational complexity. Thus, it can be concluded that the proposed method can effectively resist reverberation and noise.

Author Contributions: Conceptualization, Q.H. and W.F.; methodology, Q.H. and W.F.; software, W.F.; validation, Q.H. and W.L Fang.; formal analysis, Q.H.; investigation, W.F.; resources, Q.H.; data curation, W.F.; writing—original draft preparation, W.F.; writing—review and editing, Q.H. and W.F.; visualization, W.F.; supervision, Q.H.; project administration, Q.H.; funding acquisition, Q.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 61571279.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors received the support of the National Natural Science Foundation of China with the grant number 61571279. Sponsors provided financial support only. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Knight, W.C.; Pridham, R.G.; Kay, S.M. Digital signal processing for sonar. *Proc. IEEE* **1981**, *69*, 1451–1506. [CrossRef]
2. Boukerche, A.; Oliveira, H.A.B.F.; Nakamura, E.F.; Loureiro, A.A.F. Localization systems for wireless sensor networks. *IEEE Wirel. Commun.* **2007**, *14*, 6–12. [CrossRef]
3. Krim, H.; Viberg, M. Two decades of array signal processing research: The parametric approach. *IEEE Signal Process. Mag.* **1996**, *13*, 67–94. [CrossRef]

4. Gannot, S.; Vincent, E.; Markovich-Golan, S.; Ozerov, A. A Consolidated Perspective on Multimicrophone Speech Enhancement and Source Separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 692–730. [CrossRef]
5. Wang, H.; Chu, P. Voice source localization for automatic camera pointing system in videoconferencing. In Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, Munich, Germany, 21–24 April 1997; Volume 1, pp. 187–190.
6. Yu, Y.; Wang, W.; Luo, J.; Feng, P. Localization based stereo speech separation using deep networks. In Proceedings of the 2015 IEEE International Conference on Digital Signal Processing (DSP), Singapore, 21–24 July 2015; pp. 153–157.
7. Schmidt, R. Multiple emitter location and signal parameter estimation. *IEEE Trans. Antennas Propag.* **1986**, *34*, 276–280. [CrossRef]
8. Roy, R.; Kailath, T. ESPRIT-estimation of signal parameters via rotational invariance techniques. *IEEE Trans. Acoust. Speech Signal Process.* **1989**, *37*, 984–995. [CrossRef]
9. Capon, J. High-resolution frequency-wavenumber spectrum analysis. *Proc. IEEE* **1969**, *57*, 1408–1418. [CrossRef]
10. Yang, J.-M.; Lee, C.-H.; Kim, S.; Kang, H.-G. A robust time difference of arrival estimator in reverberant environments. In Proceedings of the 2009 17th European Signal Processing Conference, Glasgow, Scotland, 24–28 August 2009; pp. 864–868.
11. Knapp, C.; Carter, G. The generalized correlation method for estimation of time delay. *IEEE Trans. Acoust. Speech Signal Process.* **1976**, *24*, 320–327. [CrossRef]
12. Lemos, R.P.; e Silva, H.V.L.; Flores, E.L.; Kunzler, J.A.; Burgos, D.F. Spatial Filtering Based on Differential Spectrum for Improving ML DOA Estimation Performance. *IEEE Signal Process. Lett.* **2016**, *23*, 1811–1815. [CrossRef]
13. Huang, Z.; Xu, J.; Pan, J. A Regression Approach to Speech Source Localization Exploiting Deep Neural Network. In Proceedings of the 2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM), Xi'an, China, 13–16 September 2018; pp. 1–6.
14. Zhu, W.; Zhang, M. A Deep Learning Architecture for Broadband DOA Estimation. In Proceedings of the 2019 IEEE 19th International Conference on Communication Technology (ICCT), Xi'an, China, 16–19 October 2019; pp. 244–247.
15. Zhu, W.; Zhang, M.; Li, P.; Wu, C. Two-Dimensional DOA Estimation via Deep Ensemble Learning. *IEEE Access* **2020**, *8*, 124544–124552. [CrossRef]
16. Diaz-Guerra, D.; Miguel, A.; Beltran, J.R. Robust Sound Source Tracking Using SRP-PHAT and 3D Convolutional Neural Networks. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 300–311. [CrossRef]
17. Shimada, K.; Koyama, Y.; Takahashi, N.; Takahashi, S.; Mitsufuji, Y. Accdoa: Activity-Coupled Cartesian Direction of Arrival Representation for Sound Event Localization and Detection. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 915–919.
18. Xiao, X.; Zhao, S.; Zhong, X.; Jones, D.L.; Chng, E.S.; Li, H. A learning based approach to direction of arrival estimation in noisy and reverberantenvironments. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 19–24 April 2015; pp. 2814–2818.
19. Stöter, F.; Chakrabarty, S.; Edler, B.; Habets, E.A.P. Classification vs. Regression in Supervised Learning for Single Channel Speaker Count Estimation . In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 436–440.
20. Perotin, L.; Défossez, A.; Vincent, E.; Serizel, R.; Guérin, A. Regression Versus Classification for Neural Network Based Audio Source Localization. In Proceedings of the 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 20–23 October 2019; pp. 343–347.
21. Varanasi, V.; Gupta, H.; Hegde, R.M. A Deep Learning Framework for Robust DOA Estimation Using Spherical Harmonic Decomposition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 1248–1259. [CrossRef]
22. Chakrabarty, S.; Habets, E.A.P. Broadband DOA estimation using convolutional neural networks trained with noise signals. In Proceedings of the 2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 15–18 October 2017.
23. Adavanne, S.; Politis, A.; Virtanen, T. Sound Event Localization and Detection of Overlapping Sources Using Convolutional Recurrent Neural Networks. *IEEE J. Sel. Top. Signal Process.* **2018**, *13*, 34–48. [CrossRef]
24. Liu, Z.; Zhang, C.; Yu, P.S. Direction-of-Arrival Estimation Based on Deep Neural Networks with Robustness to Array Imperfections. *IEEE Trans. Antennas Propag.* **2018**, *66*, 7315–7327. [CrossRef]
25. Perotin, L.; Serizel, R.; Vincent, E.; Guérin, A. CRNN-based Joint Azimuth and Elevation Localization with the Ambisonics Intensity Vector. In Proceedings of the 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC), Tokyo, Japan, 17–20 September 2018.
26. Rafaely, B. Analysis and design of spherical microphone arrays. *IEEE Trans. Speech Audio Process.* **2005**, *13*, 135–143. [CrossRef]
27. Avargel, Y.; Cohen, I. On multiplicative transfer function approximationin the short-time fourier transform domain. *IEEE Signal Process. Lett.* **2007**, *14*, 337–340. [CrossRef]
28. Nadiri, O.; Rafaely, B. Localization of multiple speakers under high reverberation using a spherical microphone array and the direct-pathdominance test. *IEEE Trans. Audio Speech Lang. Process.* **2014**, *22*, 1494–1505. [CrossRef]
29. Rafaely, B. Plane-wave decomposition of the pressure on a sphere by spherical convolution. *J. Acoust. Soc. Am.* **2004**, *116*, 2149–2157. [CrossRef]
30. Paulraj, A.; Roy, R.; Kailath, T. Estimation of signal parameters via rotational invariance techniques-esprit. In Proceedings of the Nineteeth Asilomar Conference on Circuits, Systems and Computers, Pacific Grove, CA, USA, 6–8 November 1985; Volume 37, pp. 83–89.

31. Huang, Q.; Chen, T. One-Dimensional MUSIC-Type Algorithm for Spherical Microphone Arrays. *IEEE Access* **2020**, *8*, 28178–28187. [CrossRef]
32. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
33. Wang, C.; Xu, Q.; Li, X.; Zheng, G.; Liu, B.; Cheng, Y. An Objective Technique for Typhoon Monitoring with Satellite Infrared Imagery. In Proceedings of the 2019 Photonics & Electromagnetics Research Symposium-Fall (PIERS-Fall), Xiamen, China, 17–20 December 2019; pp. 3218–3221.
34. Singh, P.P.; Kaushik, R.; Singh, H.; Kumar, N.; Rana, P.S. Convolutional Neural Networks Based Plant Leaf Diseases Detection Scheme. In Proceedings of the 2019 IEEE Globecom Workshops (GC Wkshps), Waikoloa, HI, USA, 9–13 December 2019; pp. 1–7.
35. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the 3rd International Conference for Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015.
36. Jarrett, D.P.; Habets, E.A.P.; Thomas, M.R.P.; Naylor, P.A. Rigidsphere room impulse response simulation: Algorithm and applications. *J. Acoust. Soc. Am.* **2012**, *132*, 1462–1472. [CrossRef] [PubMed]
37. The Eigenmike Microphone Array. 2013. Available online: http://www.mhacoustics.com/ (accessed on 3 April 2022).
38. Speech activity detection and enhancement of a moving speaker based on the wideband generalized likelihood ratio and microphone arrays. *J. Acoust. Soc. Am.* **2004**, *116*, 2406–2415. [CrossRef] [PubMed]
39. Varzandeh, R.; Adiloğlu, K.; Doclo, S.; Hohmann, V. Exploiting Periodicity Features for Joint Detection and DOA Estimation of Speech Sources Using Convolutional Neural Networks. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 566–570.
40. Panayotov, V.; Chen, G.; Povey, D.; Khudanpur, S. Librispeech: An ASR corpus based on public domain audio books. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 19–24 April 2015; pp. 5206–5210.
41. Evers, C.; Löllmann, H.W.; Mellmann, H.; Schmidt, A.; Barfuss, H.; Naylor, P.A.; Kellermann, W. The LOCATA Challenge: Acoustic Source Localization and Tracking. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 1620–1643. [CrossRef]
42. Kumar, L.; Bi, G.; Hegde, R.M. The spherical harmonics root-music. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 3046–3050.
43. Rafaely, B.; Kolossa, D. Speaker localization in reverberant rooms based on direct path dominance test statistics. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 6120–6124.
44. Tang, Z.; Kanu, J.D.; Hogan, K.; Manocha, D. Regression and Classification for Direction-of-Arrival Estimation with Convolutional Recurrent Neural Networks. *arXiv* **2019**, arXiv:1904.08452.
45. Pei, S.; Chang, J.; Ding, J.; Chen, M. Eigenvalues and Singular Value Decompositions of Reduced Biquaternion Matrices. *IEEE Trans. Circuits Syst. I Regul. Pap.* **2008**, *55*, 2673–2685.
46. Rubsamen, M.; Gershman, A.B. Direction-of-arrival estimation for nonuniform sensor arrays: From manifold separation to fourier domain music methods. *IEEE Trans. Signal Process.* **2009**, *57*, 588–599. [CrossRef]
47. Cheng, Y.; Yu, F.X.; Feris, R.S.; Kumar, S.; Choudhary, A.; Chang, S.-F. An exploration of parameter redundancy in deep networks with circulant projections. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 2857–2865.