JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 38, 445-461 (2022)
DOI: 10.6688/JISE.202203_38(2).0010

Location Estimation of Receivers in an Audio Room using Deep Learning with a Convolution Neural Network*

MINH-TUAN NGUYEN[1,2] AND JIN-H. HUANG[1]
[1]*Department of Mechanical Engineering*
Feng Chia University
Taichung, 407 Taiwan
[2]*Department of Mechanical Engineering*
Hung Yen University of Technology and Education
Hung Yen, 16000 Vietnam
E-mail: tuanctm7@mail.com; jhhuang@fcu.edu.tw

The audio signal obtained by a receiver from a sound source depends on the sound environment and the location of the receiver relative to the source. When an audio signal is given, it is necessary to find the best location of the receiver to obtain the audio signal. This paper presents a sound receiver location estimation method using a convolutional neural network. The sound receiver's location estimation task is comprehended as an image classification problem; in which we aim to classify a given audio signal according to the location of the receiver. Rectangular audio rooms are simulated with different dimensions and surface materials. The audio signal obtained by a receiver from a fixed sound source in the simulation room is calculated and simulated via impulse response by the image source model. Then, the audio signals are transformed into spectrograms, allowing the convolutional layers to extract the appropriate features required for classification. After datasets are trained and tested, the proposed convolutional neural network model with optimal hyperparameters exhibits high audio signal identification accuracies for all the simulation rooms. Using the proposed model, an experiment testing the receiver's estimated location in an experiment room was conducted, and the results indicate an identification accuracy of 97.6%. The research can also be applied to obtain optimal sound quality and design of an audio room.

Keywords: sound receiver location estimation, image classification, audio signal processing, image source model, convolution neural network

1. INTRODUCTION

The intensity measurement of a fixed sound source depends on the location of the receivers. For instance, in the case of an outgoing spherical wave, the sound intensity from a sound source will be inversely proportional to the square of the distance if the medium is lossless. In addition, when the sound reaches the receiver, the directional sensitivity of the receiver and the physical presence of the receiver in the sound field may alter the intensity of the final perceived or recorded sound. Therefore, the estimation of the sound receiver's location is a necessary process that determines the location of the receiver for obtaining a given audio signal, which can also be used as a supplement to audio signal processing research.

Important applications of sound signal processing are audio data compression [1], a

Received August 11, 2020; revised November 16 & December 16, 2020; accepted December 29, 2020.
Communicated by Jen-Tzung Chien.
* This research was supported by the Ministry of Science and Technology of Taiwan under Contract Nos. MOST-107-2221-E-035-074-MY3 and MOST108-2218-E-035-007.

synthesis of audio effects [2], and audio signal classification. With audio compression becoming the most prominent application of digital audio processing, the burgeoning importance of multimedia content management has been experiencing growing applications of signal processing in audio signal segmentation [3] and classification. Audio signal classification is a part of the larger problem of audiovisual data handling. It has a lot of important applications in digital libraries, professional media production, education, entertainment, and surveillance systems. Classic problems, such as speech and speaker recognition, have been widely considered for decades. There are several studies in the field of audio signal classification, such as musical genre classification [4-7], musical instrument recognition [8, 9], speaker recognition [10, 11], language recognition [12-14], audio context recognition [15], video segmenting based on audio [16], and sound effects retrieval [17]. In addition, audio signal classification can be applied to estimate the relative location between a sound source and receiver. Most research has focused on sound source localization, such as the azimuth aspect only [18-20], the azimuth and elevation aspects only [21, 22], and the distance aspect only [23]. Studies that investigate the location of the receiver are relatively few in number [24, 25]. In fact, many sound systems have fixed sound sources. In this case, the obtained audio signal in a sound system depends only on the audio environment and the location of the receiver. Therefore, it is necessary to estimate the location of the receiver to obtain the desired audio signal, as well as the optimization of the sound quality.

Many studies have been conducted in the past few years regarding the methods and techniques applied in audio signal processing. These studies have focused on audio signal classification and segmentation using several features and techniques. In 2005, Lin *et al.* [26] implemented a bottom-up support vector machine on acoustic features such as subband power, pitch information, and additional parameters, such as frequency cepstral coefficients, to accomplish audio classification and categorization. Audio feature extraction and a multigroup classification scheme focusing on identifying discriminatory time-frequency subspaces using the local discriminant bases technique was proposed by Umapathy *et al.* [27] in 2007. Based on the calculated features, such as linear prediction coefficients and linear prediction cepstral coefficients, a clustering algorithm was applied to structure the music content by Xu *et al.* [28]. Ajmera *et al.* [29] provided an approach that uses an artificial neural network and hidden Markov model toward high-performance speech/music discrimination on realistic tasks related to the automatic transcription of broadcast news. A method was proposed for speech/music discrimination based on root mean square and zero-crossings, as described in [30]. Another method was proposed by Honda *et al.* [31] for estimating the distance of single-channel audio signals; here, the signal's distance was estimated by phase interference between the observed and pseudo-observed signal waves.

Researchers have relied on handcrafted features to estimate sound source distances in the previous studies of the relative location between the sound source and the receiver. A method for distance perception in rooms that applies the information in the room impulse response was proposed by Bronkhorst *et al.* [32]. Lu *et al.* [33] suggested a binaural distance estimation method using the direct-to-reverberant-ratio, which was accomplished by first estimating the sound source's direction and then removing the energy of the sound in that region to identify the reverberant signal. Rodemann *et al.* [34] estimated sound source distances using several audio cues, including the interaural intensity difference, the interaural temporal difference, sound amplitude, and spectral characteristics, discovering

that in certain circumstances, mean signal amplitude and binaural cues can provide a very reliable distance estimation. However, using handcrafted features to estimate distances may make the extraction process complicated or faint. In addition, the accuracy of these methods is far from perfect and needs to be improved.

In recent years, deep neural networks have been widely used for sound source localization methods. Huang *et al.* [35] successfully determined the range of sound pressure by applying deep neural networks. A convolutional recurrent neural network was proposed by Yiwere *et al.* [23] to estimate the sound source distances in known environments. Yalta *et al.* [36] proposed the use of a deep neural network to localize a sound source using an array of microphones in a reverberant environment. Takeda *et al.* [37] solved sound source localization based on deep neural networks using discriminative training; the results showed the remarkable performance of a deep learning model in audio signal processing in general and sound source localization in particular. Through the current literature survey of sound source localization, it was found that the challenges in sound source localization are how to identify the suitable features of audio signals and improve accuracy.

In the current study, three rectangular audio rooms were created based on the image source model (ISM) [38]. The shape, dimensions, and materials of the audio room surfaces are simulated. A single sound source is given a fixed location in the room, and the audio signal emitted from this source is recorded by a receiver. The received audio signal has then extracted a feature as a spectrogram using Short-time Fourier transform (STFT). A convolution neural network (CNN) is applied to estimate the location of the receiver as an audio signal classification problem from the input images, which are the spectrograms. The transformation of audio signals into the spectrograms eliminates the need for complex handcrafted techniques to extract the features of the audio signal, allowing the convolution layers to extract the appropriate features automatically for the classification. The results show that the classification accuracy is very high, over 97% in simulation and experiment.

2. METHODOLOGY

2.1 Main Framework

The main framework of the sound receiver's location estimation is divided into two phases: the training phase and the test phase, as shown in Fig. 1. In the training phase, the data are inserted in the form of signal and label, and then, a segmentation technique is applied to the training data, followed by a feature extraction technique based on the spectrogram. At last, the CNN model is trained. In the testing phase, the data are inserted as a signal. After undergoing segmentation, the prediction results for the location of the receiver are obtained by the trained CNN model. The accuracy of the prediction model would be evaluated using the evaluation parameter.

2.2 Convolution Neural Network

A CNN is a deep learning algorithm that by the organization and functionality of the visual cortex and is designed to mimic the connectivity pattern of neurons within the human brain [39]. It can be used to recognize and classify features in computer vision and audio recognition. A CNN is a multilayer neural network designed to analyze visual inputs and perform tasks such as image classification, segmentation, and object detection, which

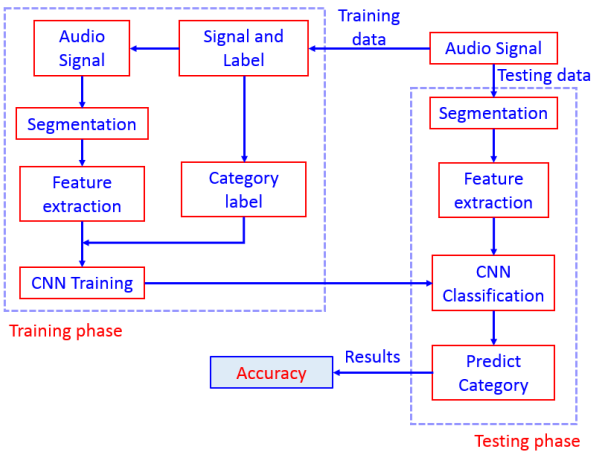


Fig. 1. The main framework of the sound receiver's location estimation.

might be useful for autonomous vehicles. A CNN can also be used for deep learning applications in acoustic signal processing such as speech recognition and sound classification. A CNN consists of two main parts: a convolution layer and a fully connected layer. The convolution layer splits various features of the image for the analysis, while the fully connected layer uses the output of the convolution layer to predict the most effective description for the image.

In the current study, a classification model is designed with reference to the models in previous studies [23, 36] and depicted in Fig. 2. The classification model includes two convolution layers that consist of a set of learnable filters and one fully-connected layer. The first convolution layer contains 16 3×3 pixels size filters; the second convolution layer contains 32 3×3 pixels size filters. In a convolution layer, a 2D convolutional layer first extracts features from the input image and then preserves the relationship between the pixels by learning image features using small squares of an input image. The filter slides vertically and horizontally along the input image-guided by the number of pixels the filter moves on at a time called the stride. The ReLU activation function is applied to each convolution layer's output, and the activation maps are max-pooled to reduce their dimensions. Max-pooling takes the largest element from the feature map using a 5×5 pixels window. The ReLU activation function and Max-Pooling algorithm are illustrated in Eq. (1) and Fig. 3, respectively. After max-pooling, the feature maps from the convolution layer are reshaped. Next, the CNN layer's output is passed to the fully connected layer, containing

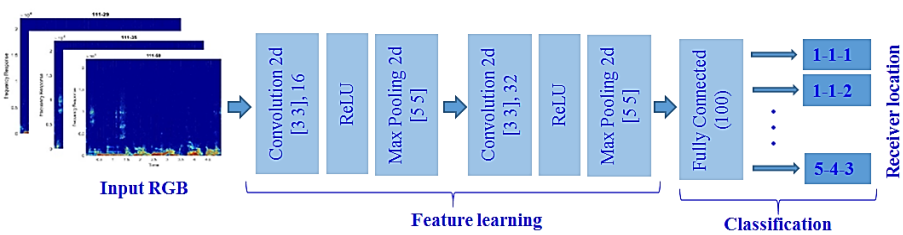


Fig. 2. Description of the proposed CNN architecture.

Fig. 3. Max-Pooling algorithm using a 2×2 pixels window and a stride of 2 pixels.

100 neurons. Finally, an activation function is the softmax function that is used to classify the outputs. Mathematically, the softmax function is expressed as Eq. (2), where $\boldsymbol{x}$ is the input vector, and $j(= 1, 2, \ldots, N)$ denotes the output unit.

$$ReLU(s) = \max(s, 0) \tag{1}$$

$$softmax(\boldsymbol{x}_j) = \frac{e^{x_j}}{\sum_{n=1}^{N} e^{x_n}} \tag{2}$$

The loss function the network used for training is the mean squares error (MSE), the average of the squared difference between the true label and prediction labels. The equation of the loss function is defined as follows:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(Y_{true}(x(i)) - Y_{predict}(x(i)))^2 \tag{3}$$

where N is the number of examples in the dataset and $Y_{true}(x(i))$ and are the true and $Y_{predict}(x(i))$ prediction labels of example $x(i)$, respectively.

The CNN model with setting parameters was used to classify audio signals according to the receiver's location in three simulation rooms. A set of appropriate hyperparameters was found by experience and used in the training process to ensure a stable training process and high accuracy. This model, with the same set of hyperparameters, was also used for an experiment room.

3. SIMULATION

3.1 Simulation Rooms

As displayed in Fig. 4, the simulation room is modeled as a three-dimensional space bounded by six rectangular faces with dimensions $L_x \times L_y \times L_z$. The room also contains a sound source and an acoustic receiver. The sound source is located at $p_s = [x_s\ y_s\ z_s]^T$, and the acoustic receiver is located at $p_r = [x_r\ y_r\ z_r]^T$ with direction $d_r = [a_r\ e_r]^T$, where $[x_s\ y_s\ z_s]$ and $[x_r\ y_r\ z_r]$ are the corresponding position coordinates of the source and receiver in the room, and $[a_r\ e_r]$ is the receiver's azimuth and elevation.

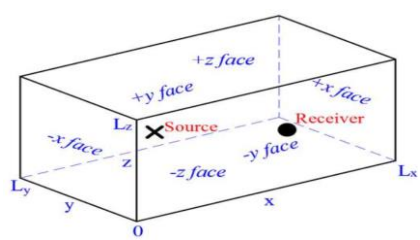


Fig. 4. Configuration of the simulation room.

In the current study, three rectangular simulation rooms were generated with different dimensions and face materials based on ISM; they are denoted as Rooms A, B, and C, respectively. Room A is the largest with dimensions of 20m×10m×5m, like a hall; Room B is 6m×5m×4m, like a classroom; Room C is 4m×3m×3m, representing a small discussion room. The dimensions, materials, location of the sound source, and receivers of these simulation rooms are listed in Table 1.

Table 1. Dimensions, face materials, the sound source's location, and the receiver's location of the simulation rooms.

Room A	Room dimensions	$L_x \times L_y \times L_z$ = 20m × 10m × 5m; 1000m^3 volume					
	Face material	$-z$	$-y$	$-x$	$+x$	$+y$	$+z$
		Plywood	Concrete block	Concrete block	Window glass	Concrete block	Plywood
	Source's location	(x_s, y_s, z_s) = (3.0m, 5.0m, 1.5m)					
	Receiver's direction	(a_r, e_r) = (0°, 0°)					
Room B	Room dimensions	$L_x \times L_y \times L_z$ = 6.0m × 5.0m × 4.0m; 120m^3 volume					
	Face material	$-z$	$-y$	$-x$	$+x$	$+y$	$+z$
		Platform wood	Concrete block painted	Brick	Draperies	Concrete block painted	Plaster sprayed
	Source's location	(x_s, y_s, z_s) = (2.0m, 2.5m, 3.0m)					
	Receiver's direction	(a_r, e_r) = (90°, 180°)					
Room C	Room dimensions	$L_x \times L_y \times L_z$ = 4.0m × 3.0m × 3.0m; 36 m^3 volume					
	Face material	$-z$	$-y$	$-x$	$+x$	$+y$	$+z$
		Carpet on felt	Plaster on lath	Plaster on lath	Window glass	Plaster on lath	Plaster sprayed
	Source location	(x_s, y_s, z_s) = (0.5m, 0.5m, 2.5m)					
	Receiver's direction	(a_r, e_r) = (–90°, 90°)					
Sample frequency f_s = 44,100Hz							

The faces are described by frequency-dependent absorption coefficients that can be selected from Hall [40]. Table 2 lists the absorption coefficients of the face materials for each frequency band.

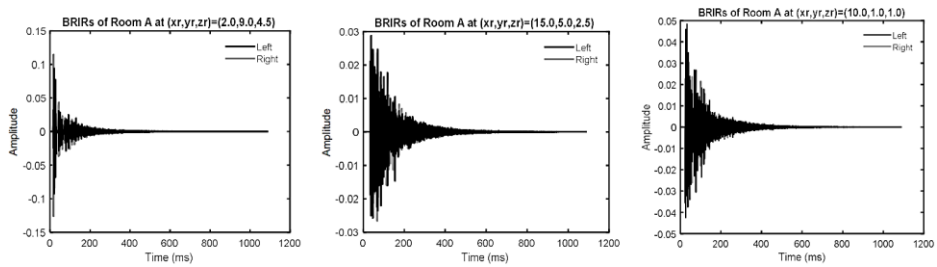
From the modeled simulation rooms with the location of the given sound source and receiver, the binaural room impulse responses (BRIRs) and audio signal can be obtained. Fig. 5 illustrates the BRIR results of the receiver at three different locations in Rooms A, B, and C.

Fig. 5 indicates that in the same room, the received audio signal depends on the receiver's location. For example, in Room A, three received audio signals corresponding to the distance from the receiver to the source are 12.04m, 8.08m, and 5.10m, and their amplitudes are different. When the location of the receiver is far from the source, the amplitude of the obtained sound is smaller. This is because the sound travels farther. In addition, the sound intensity is also reduced because of the absorption of air and its impinging sur-

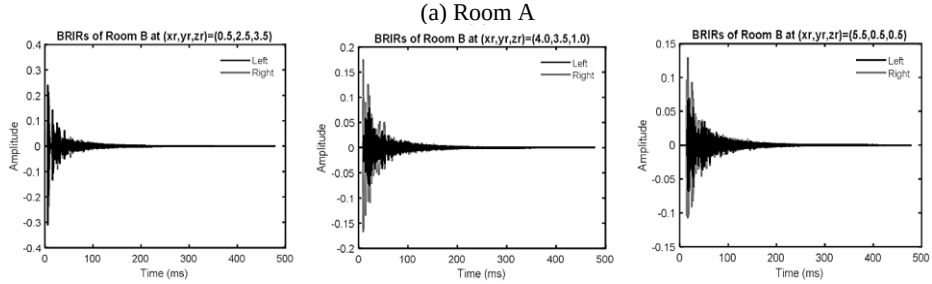
faces. However, the reverberation time of the audio signal remains the same. The sound intensity and propagation time also depend on the room's size. The larger the room, the smaller the amplitude of sound and the longer the reverberation time. Among the three simulation rooms, in the largest, Room A, the amplitude of the obtained sound is the smallest, and the reverberation time is the longest. On the other hand, the smallest one, Room C, has the largest sound amplitude and the shortest reverberation time.

Table 2. Absorption coefficients of face materials depend on the frequency band.

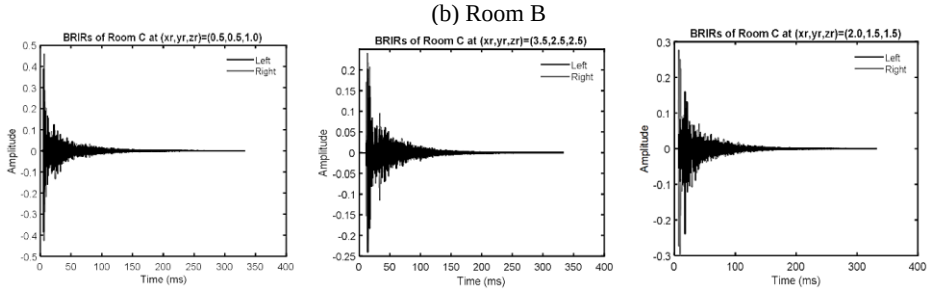
Frequency band (Hz)	125	250	500	1000	2000	4000
	Absorption coefficient					
Ply-wood	0.60	0.30	0.10	0.10	0.10	0.10
Platform wood	0.40	0.30	0.20	0.20	0.15	0.10
Concrete block	0.40	0.40	0.30	0.30	0.40	0.30
Concrete block painted	0.10	0.05	0.06	0.07	0.10	0.10
Plaster sprayed	0.50	0.70	0.60	0.70	0.70	0.50
Plaster on lath	0.20	0.15	0.10	0.05	0.04	0.05
Window glass	0.30	0.20	0.20	0.10	0.07	0.04
Brick	0.03	0.03	0.03	0.04	0.05	0.07
Draperies	0.07	0.30	0.50	0.70	0.70	0.60
Carpet on felt	0.10	0.30	0.40	0.50	0.60	0.70



(a) Room A



(b) Room B



(c) Room C

Fig. 5. BRIRs results of the receiver at the three different locations.

3.2 Data Collection

Fig. 6 demonstrates that the simulation rooms were divided into $m \times n \times k$ smaller rectangulars as the classes of the receiver's location in the room. Fifty files of audio signals in wav format are collected in each class corresponding to 50 random receiver's locations. Therefore, the dataset of each simulation room is $50 \times m \times n \times k$ audio signals. Table 3 shows the number of classes and audio signals corresponding to the three simulation rooms.

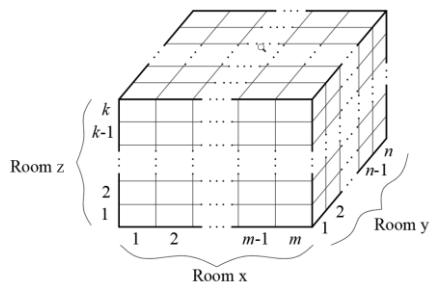


Fig. 6. Receiver's location division classes in the simulation rooms.

Table 3. The number of classes and audio signals for each simulation room.

Simulation room	Number of classes	Number of audio signals
Room A	5×4×3=60	3,000
Room B	4×3×3=36	1,800
Room C	3×3×3=27	1,350

3.3 Feature Extraction

The audio signals were divided into ten segments of 5-second signal with 50% overlap and categorized the corresponding labels, then analyzed using STFT [41]. The STFT of a discrete-time signal $x(n)$ with angular frequency ω is defined as

$$X_w(mL, \omega) = \sum_{n=-\infty}^{\infty} x(n) w(mL - n) e^{-j\omega n} \tag{4}$$

where the subscript w in $X_w(mL, \omega)$ denotes the analysis window $w(n)$. The parameter L is an integer that denotes the separation in time between adjacent short-time sections. For a fixed value of m, $X_w(mL, \omega)$ represents the Fourier transform with respect to n of the short-time section $f_m(n) = x(n)w(mL - n)$.

The number of frequency points is used to calculate the discrete Fourier transforms (DFT) is equal to the larger of 256 or the next power of two greater than the segment length. The visual representation of the STFT is a spectrogram. The spectrogram of the STFT is expressed as

$$s(x(n)) = 20 \log_{10} \left| \frac{X_w(mL, \omega)}{2*10^5} \right| / 100. \tag{5}$$

The feature extraction is done by setting a threshold and delete the data below the threshold to derive more outstanding features from the spectrogram. The threshold setting eliminates small-amplitude values that may cause interference while filtering the noise,

highlighting the signal's features. Assuming the simulation rooms and the experiment room are similar to office conditions, the noise's amplitude ranges from 40dB to 60 dB [42]; we tested some threshold values such as 40dB and 50 60dB. It was resulting in a 50dB value for high and stable accuracy. Therefore, it is selected as the threshold value in this study. The thresholding algorithm is expressed as

$$s = \begin{cases} s, & s \geq s_0 \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where s is computed as the amplitude and s_0 = 50dB is the threshold value.

Fig. 7 depicts the spectrogram of three simulation rooms with and without a threshold. The figure shows that the spectrograms applying the thresholding algorithm have better performance. This will lead to an improvement in accuracy for the CNN model.

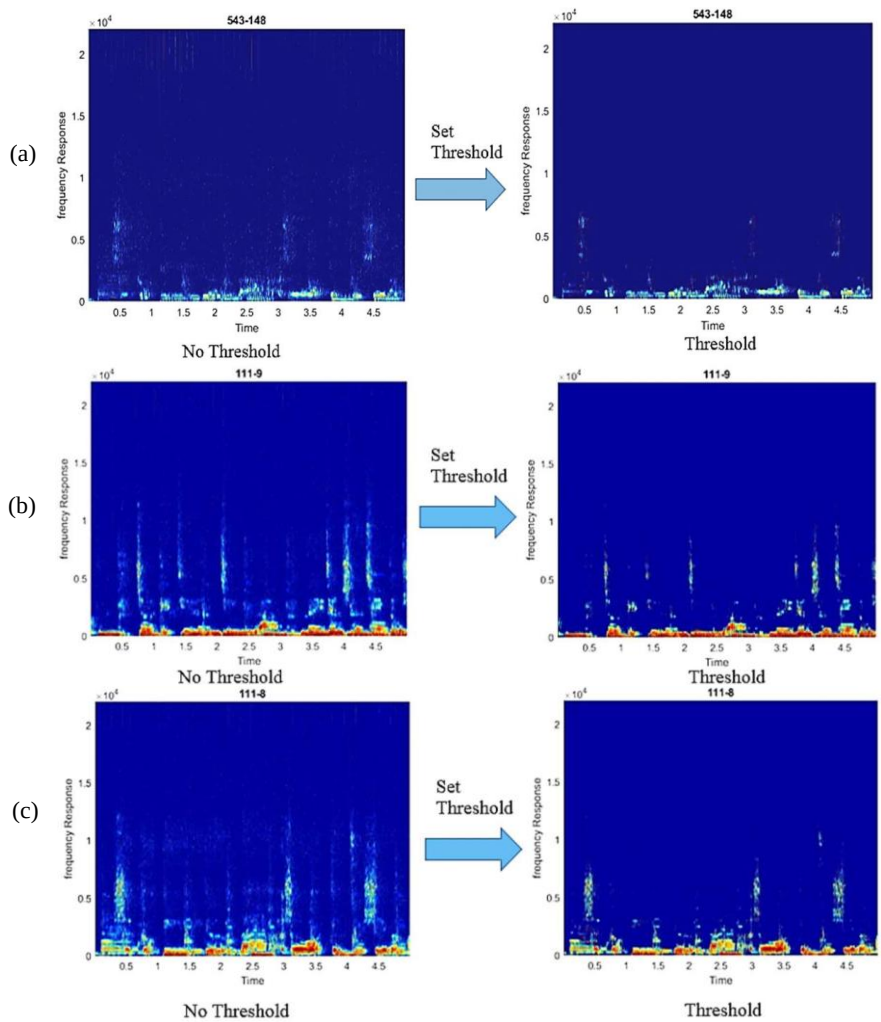
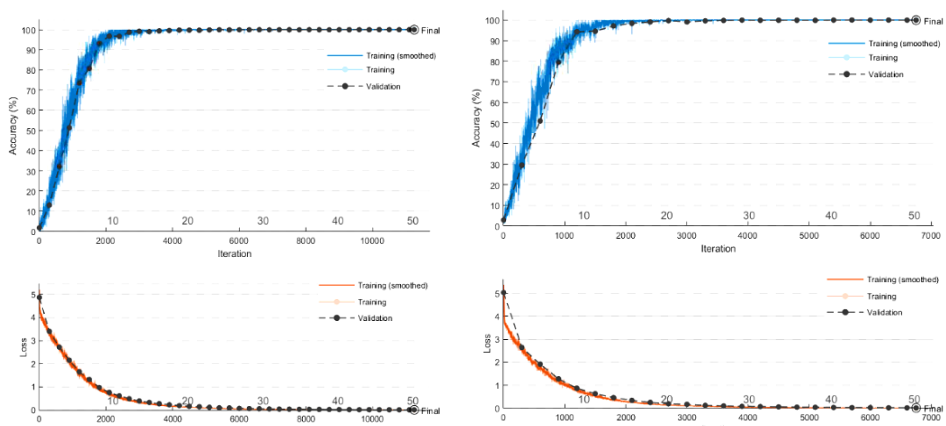


Fig. 7. Spectrogram without and with the threshold in the feature extraction of an audio signal in (a) Room A; (b) Room B; (c) Room C.

3.4 Simulation Results and Discussion

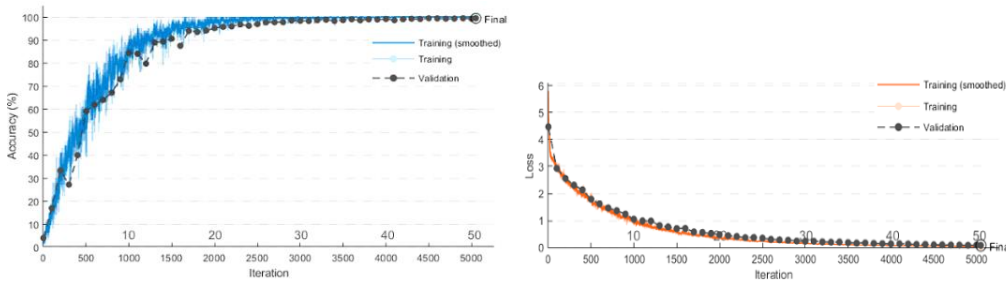
Input images of a dataset with a resolution of 224×224×3 pixels were split into a training set and a testing set. The training set contains 75% of the whole dataset, and the testing set contains 25% of the whole dataset. The optimizer is the adaptive moment estimation (Adam). Some training tests were performed to find the set of appropriate hyperparameters that ensure high accuracy and a stable training process. As a result, a set of appropriate hyperparameters was found. Specifically, the learning rate is 10^{-6}, the number of epochs is 50, and the batch size is 100. The CNNs were programmed, implemented on MATLAB R2019b software using an Intel Core i9-9000K 3.6GHz CPU on a computer equipped with NVIDIA GTX 1080Ti GPU.

For each simulation room, the input dataset includes $500 \times m \times n \times k$ images divided into $m \times n \times k$ classes corresponding to the receiver's location. The name of these classes is denoted by the corresponding location index, from "1-1-1" to "*m*-*n*-*k*" in Room x, Room y, and Room z directions. The number of input images, the number of classes, and the training progresses and results of the simulation rooms are shown in Fig. 8 and Table 4. This result shows that the audio signal identification accuracies are 99.9%, 99.9%, and 99.4%, respectively, for Rooms A, B, and C. The accuracy and loss curves show the stability of the training process in these three rooms. This stability is also reflected in the confusion matrix of Room C, as shown in Fig. 9. This figure demonstrates the confusion matrix for each classification task with the highest accuracy. In the confusion matrix, each row represents the instances in a true class and each column represents the instances in a predicted class. The correctly predicted instances are shown in the diagonal of the matrix, while the values outside the diagonal show the incorrectly predicted instances. For Room C, the average accuracy is 99.4%, and the accuracy for each class is very close. Among the 27 classes, 21 classes have an accuracy of 100%, and the lowest accuracy is 95.2% of the 2-2-2 classes. Furthermore, the adjacent classes are also more closely related. For example, there are four confusion samples among 125 samples in class 1-2-3, all of which were confused into an adjacent class is 1-2-2. Similarly, there are four confusion samples in class 2-2-2, in which, two samples were confused into class 2-1-2 and the remaining two samples were confused into class 2-2-1. This is because, at the receiver's locations close to each other, the received audio signals are also close.



(a) Room A (b) Room B

Fig. 8. Accuracy and loss curves of training progress.



(c) Room C

Fig. 8. (Cont'd) Accuracy and loss curves of training progress.

Table 4. Accuracy and training time of the simulation rooms.

Simulation room	Number of input images	Number of classes	Accuracy (%)	Training time
Room A	30000	60	99.9	103 min 29 sec
Room B	18000	36	99.9	60 min 28 sec
Room C	13500	27	99.4	50 min 12 sec

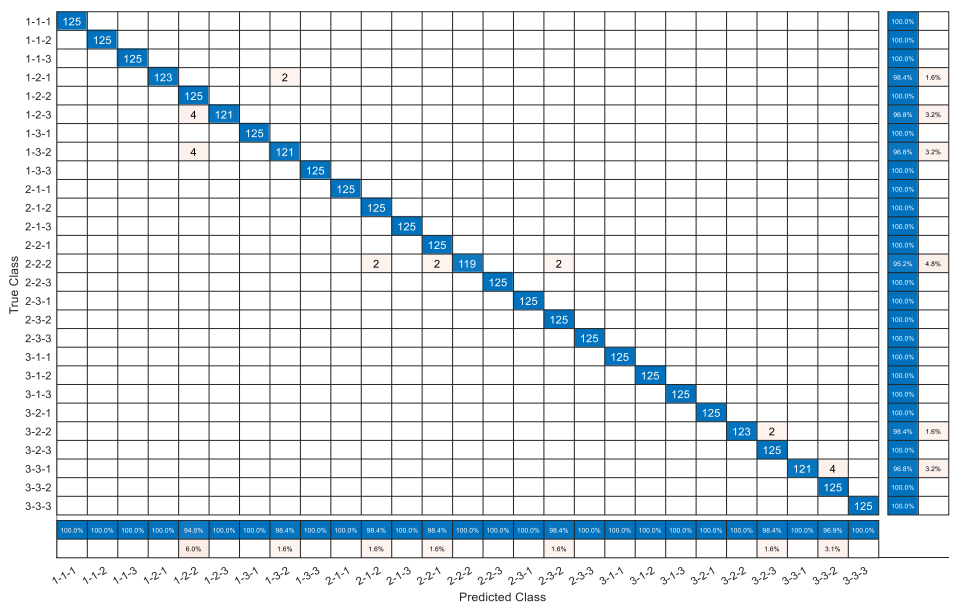


Fig. 9. Confusion matrix of Room C.

Regarding the training time, because having the largest amount of input data and classes, the training time for Room A is the longest, totaling 103 minutes 29 seconds. The value for Room C is the fastest, taking 50 minutes 12 seconds. This shows that our model with the set of appropriate hyperparameters achieves almost perfect stability and accuracy when estimating the receiver locations in all three simulation rooms. Next, the model will be used in experiments.

4. EXPERIMENT

4.1 Experiment Setup

The experiment was conducted in a facility at Feng Chia University, Taiwan. The sound source is a loudspeaker, and the receiver is a Zoom H6 handy recorder using an XY microphone with rotating mics of 120°. Fig. 10 shows the experiment room with the sound source and receiver. The dimensions and face materials of the experiment room, as well as the location of the sound source and receiver, are described in Table 5. The audio signals from the source were recorded and divided into 18 classes corresponding to the receiver's locations defined as class 1-1-1 to class 3-3-2.

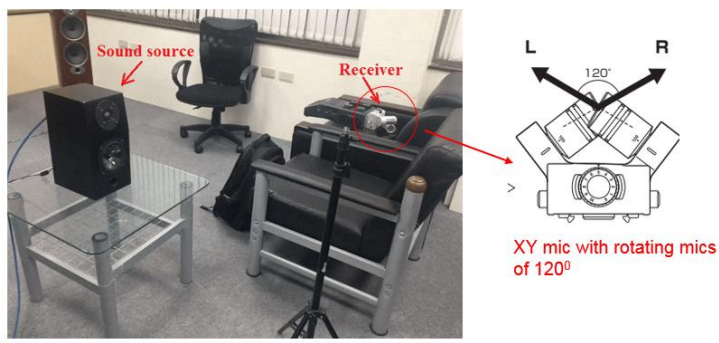


Fig. 10. Experiment room with the sound source and receiver.

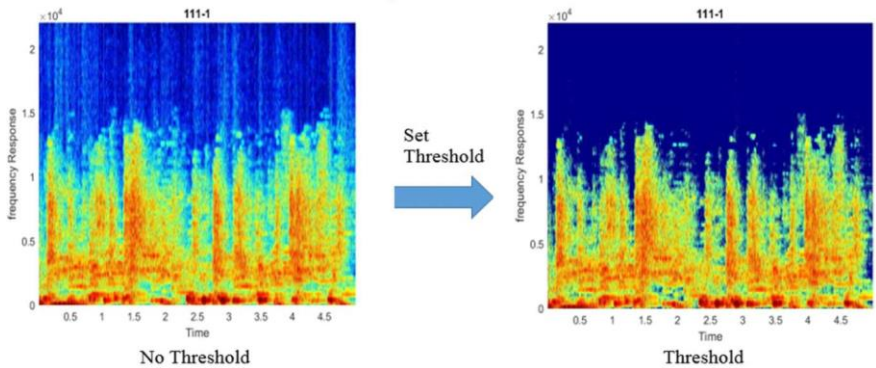
Table 5. Parameters of the experiment room.

Room dimensions	$L_x \times L_y \times L_z$ = 7.0m × 5.0m × 2.8m; 98 m^3 volume					
Face material	$-z$	$-y$	$-x$	$+x$	$+y$	$+z$
	Felt lining Plywood	Window glass	Felt lining Concrete	Platform wood	Window glass	Ply-wood
Source's location	(x_s, y_s, z_s) = (2.5m, 2.5m, 0.7m)					
Receiver's direction	(a_r, e_r) = (180°, 0°)					

4.2 Experiment Results and Discussion

The input dataset includes 5,148 segments of 5-second audio divided into 18 (3×3×2) classes corresponding to the receiver's locations from classes 1-1-1 to 3-3-2. The audio signal $x(i)$ received in a room has a sampling frequency of 44,100 Hz. The spectrogram is used for feature extraction for classification. The spectrograms of an audio signal with and without a threshold are shown in Fig. 11.

The proposed CNN model was used for training. The training set accounts for 75% of the entire dataset, and the testing set accounts for 25% of the whole dataset. Fig. 12 shows the training trend, the accuracy curve, and the loss curve during training progress before adjusting the hyperparameters.



(a) Without threshold. (b) With threshold.

Fig. 11. Spectrogram of an audio signal of the experiment room.

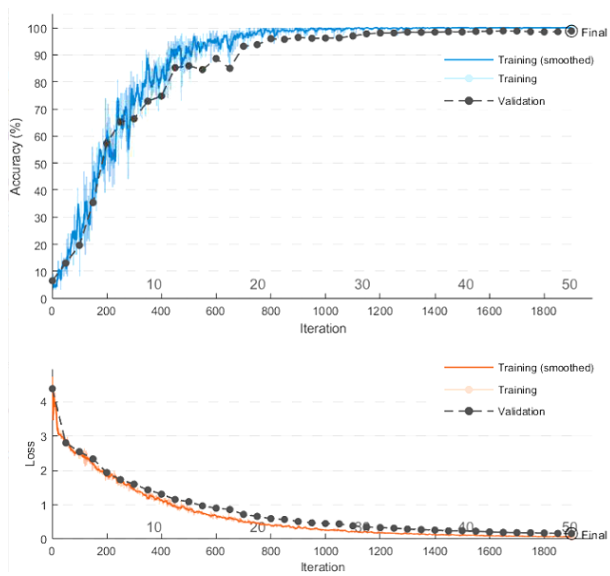


Fig. 12. Accuracy and loss curves of the training progress of the experiment room.

Fig. 12 shows the audio signal identification accuracy is 97.6%, and the accuracy and loss curves are smooth, indicating the stability of the training process. From the confusion matrix in Fig. 13, we can see that the accuracy of each class is relatively close to each other. There are nine classes among 18 with 100% accuracy, while the lowest accuracy is 97.2%. Similar to the simulation rooms, the adjacent classes are also more closely related. For example, there are two confusion samples among 71 in class 2-1-1, all of which were confused into an adjacent class is 3-1-1; in class 2-2-2, there is one confusion sample, this was confused into class 2-2-1.

Compare this result with a previous study, reported by Takeda *et al.* [37]; both studies used deep neural networks and STFT in estimating the relative location between the sound source and the receiver; our method gave significantly higher accuracy, 97.6% vs. 89.3%.

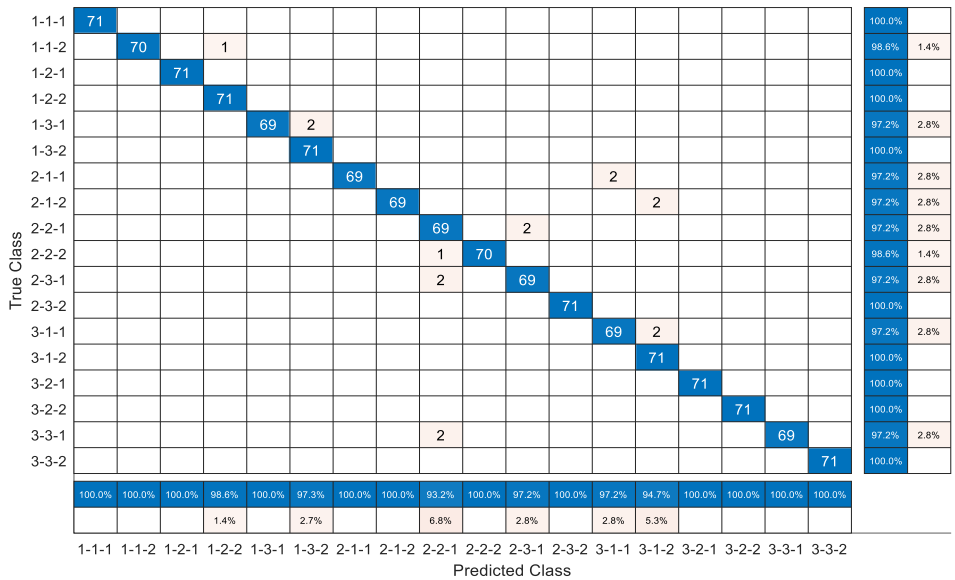


Fig. 13. Confusion matrix of the experiment room.

This result once again verifies that the proposed CNN model with the adjusted hyperparameters is a suitable model for classifying sound with the feature extraction as the spectrogram of the audio signal. Therefore, the receiver's location in the audio room was estimated with high accuracy.

5. CONCLUSION

In the current research, a method was proposed to estimate the location of a receiver in an audio room by using the designed CNN. Three rectangular audio simulation rooms were created according to the ISM, each with different sizes and different surface materials. In the simulation rooms, the audio signal emitted from a fixed sound source was recorded by the receiver at different locations. Then, the CNN model with hyperparameters that was chosen by our experience was applied to estimate the location of the receiver in the audio room. The CNN model consists of two convolution layers: a fully connected layer and an output layer. The class identification performance under the three simulation setups with 60, 36, and 27 classes, can achieve audio signal identification accuracy of 99.9%, 99.99%, and 99.4%, respectively. In addition, the experiment was conducted in a room, and the receiver's location was divided into 18 classes. Using the designed CNN model, the results of the experiment also achieved a very high identification accuracy of 97.6%.

The following phenomena are also observed: (a) In an audio room, the received audio signals at different receiver locations are different. A signal that is further from the source has a smaller amplitude than a signal that is nearer to the source. However, the reverberation time remains the same. Based on the differences, these audio signals can be classified according to the receiver's location; (b) Using a CNN to estimate the location of the receiver in an audio room with the feature extraction of the audio signal is a spectrogram that gives very high accuracy.

The results could be helpful for estimating a sound receiver's location in an audio system, thereby optimizing sound system design. Future research will focus on estimating the location of the receiver in more complex sound environments (such as multisource, multiroom) with the help of other deep learning techniques such as a recurrent neural network (RNN).

REFERENCES

1. D.-L. Zeng, *et al.*, "Audio data compression based on AVS-P10," in *Proceedings of IEEE 29th Chinese Control And Decision Conference*, 2017, pp. 3608-3612.
2. D. A. D'Souza and V. V. D. Shastrimath, "Modelling of audio effects for vocal and music synthesis in real time," in *Proceedings of IEEE 3rd International Conference on Computing Methodologies and Communication*, 2019, pp. 1-4.
3. R. H. Zottesso, *et al.*, "Automatic segmentation of audio signal in bird species identification," in *Proceedings of IEEE 35th International Conference of the Chilean Computer Science Society*, 2016, pp. 1-11.
4. G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, Vol. 10, 2002, pp. 293-302.
5. C. Xu, *et al.*, "Musical genre classification using support vector machines," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2003, pp. V-429.
6. G.-Y. Son and S. Kwon, "Classification of heart sound signal using multiple features," *Applied Sciences*, Vol. 8, 2018, p. 2344.
7. H. Kon and H. Koike, "Estimation of late reverberation characteristics from a single two-dimensional environmental image using convolutional neural networks," *Journal of the Audio Engineering Society*, Vol. 67, 2019, pp. 540-548.
8. A. Eronen and A. Klapuri, "Musical instrument recognition using cepstral coefficients and temporal features," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, 2000, pp. II753-II756.
9. S. Essid, G. Richard, and B. David, "Musical instrument recognition by pairwise classification strategies," *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, 2006, pp. 1401-1412.
10. S. Safavi, *et al.*, "Speaker recognition for children's speech," *arXiv Preprint*, 2016, arXiv:1609.07498.
11. M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in *Proceedings of IEEE Spoken Language Technology Workshop*, 2018, pp. 1021-1028.
12. W. M. Campbell, *et al.*, "Support vector machines for speaker and language recognition," *Computer Speech & Language*, Vol. 20, 2006, pp. 210-229.
13. N. Dehak, *et al.*, "Language recognition via i-vectors and dimensionality reduction," in *Proceedings of the 12th Annual Conference of International Speech Communication Association*, 2011, pp. 857-860.
14. H.-D. Yang, "Sign language recognition with the kinect sensor based on conditional random fields," *Sensors*, Vol. 15, 2015, pp. 135-147.
15. T. Heittola, *et al.*, "Audio context recognition using audio event histograms," in *Proceedings of IEEE 18th European Signal Processing Conference*, 2010, pp. 1272-1276.

16. J. S. Boreczky and L. D. Wilcox, "A hidden Markov model framework for video segmentation using audio and image features," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 1998, pp. 3741-3744.
17. T. Zhang and C.-C. Kuo, "Classification and retrieval of sound effects in audiovisual data management," in *Proceedings of IEEE Conference Record of the 33rd Asilomar Conference on Signals, Systems, and Computers*, 1999, pp. 730-734.
18. L. Wang and A. Cavallaro, "Time-frequency processing for sound source localization from a micro aerial vehicle," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 496-500.
19. S. Hakrabarty and E. A. Habets, "Broadband DOA estimation using convolutional neural networks trained with noise signals," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2017, pp. 136-140.
20. S. Chakrabarty and E. A. Habets, "Multi-speaker localization using convolutional neural network trained with noise," *arXiv Preprint*, 2017, arXiv:1712.04276.
21. T. Rodemann, *et al.*, "Using binaural and spectral cues for azimuth and elevation localization, " in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2008, pp. 2185-2190.
22. L. Perotin, *et al.*, "CRNN-based joint azimuth and elevation localization with the Ambisonics intensity vector," in *Proceedings of IEEE 16th International Workshop on Acoustic Signal Enhancement*, 2018, pp. 241-245.
23. M. Yiwere and E. J. Rhee, "Sound source distance estimation using deep learning: An image classification approach," *Sensors*, Vol. 20, 2020, p. 172.
24. N. D. Gaubitch, W. B. Kleijn, and R. Heusdens, "Auto-localization in ad-hoc microphone arrays, " in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 106-110.
25. R. Parhizkar, I. Dokmanić, and M. Vetterli, "Single-channel indoor microphone localization, " in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 1434-1438.
26. C.-C. Lin, *et al.*, "Audio classification and categorization based on wavelets and support vector machine," *IEEE Transactions on Speech and Audio Processing*, Vol. 13, 2005, pp. 644-651.
27. K. Umapathy, S. Krishnan, and R. K. Rao, "Audio signal feature extraction and classification using local discriminant bases," *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 15, 2007, pp. 1236-1246.
28. C. Xu, N. C. Maddage, and X. Shao, "Automatic music classification and summarization," *IEEE Transactions on Speech and Audio Processing*, Vol. 13, 2005, pp. 441-450.
29. J. Ajmera, I. McCowan, and H. Bourlard, "Speech/music segmentation using entropy and dynamism features in a HMM classification framework," *Speech Communication*, Vol. 40, 2003, pp. 351-363.
30. C. Panagiotakis and G. Tziritas, "A speech/music discriminator based on RMS and zero-crossings," *IEEE Transactions on multimedia*, Vol. 7, 2005, pp. 155-166.
31. S. Honda, *et al.*, "Estimating the distance to a sound source using single-channel cross-spectral method between observed and pseudo-observed waves based on Phase Interference," in *Proceedings of the 23rd International Congress on Sound & Vibration*, 2016, pp. 10-14.
32. A. W. Bronkhorst, "Modeling auditory distance perception in rooms," in *Proceedings*

of AAE Forum Acusticum, 2002, pp. 1-6.
33. Y.-C. Lu and M. Cooke, "Binaural estimation of sound source distance via the direct-to-reverberant energy ratio for static and moving sources," *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 18, 2010, pp. 1793-1805.
34. T. Rodemann, "A study on distance estimation in binaural sound localization," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010, pp. 425-430.
35. Z. Huang, *et al*., "Multiple source localization in a shallow water waveguide exploiting subarray beamforming and deep neural networks," *Sensors*, Vol. 19, 2019, p. 4768.
36. N. Yalta, K. Nakadai, and T. Ogata, "Sound source localization using deep learning models," *Journal of Robotics and Mechatronics*, Vol. 29, 2017, pp. 37-48.
37. R. Takeda and K. Komatani, "Sound source localization based on deep neural networks with directional activate function exploiting phase information," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 405-409.
38. J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, Vol. 65, 1979, pp. 943-950.
39. D. H. Hubel and T. N. Wiesel, "Receptive fields of single neurones in the cat's striate cortex," *The Journal of Physiology*, Vol. 148, 1959, p. 574.
40. D. E. Hall, *Basic Acoustics*, Wiley, 1987.
41. S. Nawab, T. Quatieri, and J. Lim, "Signal reconstruction from short-time Fourier transform magnitude," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 31, 1983, pp. 986-998.
42. Office Noise and Acoustics, https://canadasafetycouncil.org/office-noise-and-acoustics/, 2020.

Minh-Tuan Nguyen received his B.E. and M.E. degrees in Mechanical Engineering from Hanoi University of Science and Technology, Hanoi, Vietnam in 2008 and 2013. At present, he is pursuing a Ph.D. degree in Mechanical Engineering at Feng Chia University, Taiwan. His research interest includes signal processing and electroacoustics.

Jin-H. Huang has earned a Ph.D. degree in Mechanical Engineering from Northwestern University in 1992. He worked for the Department of Mechanical Engineering of the Feng Chia University from 1993 till now. Presently, he is also working as Vice President, FCU. His research interest is in the areas of electroacoustics, MEMS Transducers, sound quality, and acoustics of fluid-structure interactions. He is using B&K Pulse, Sound Check, and Klippel measurements since 2000 for research and education in so-und-structure interactions and electroacoustic engineering analysis.