Applied Physics Letters ARTICLE scitation.org/journal/apl

A single feature for human activity recognition using two-dimensional acoustic array

Cite as: Appl. Phys. Lett. **114**, 214101 (2019); doi: 10.1063/1.5096572
Submitted: 19 March 2019 · Accepted: 23 April 2019 ·
Published Online: 28 May 2019

Xinhua Guo,[1,2,a)] Rongcheng Su,[1] Chaoyue Hu,[1] Xiaodong Ye,[1] Huachun Wu,[1,2] and Kentaro Nakamura[3,a),b)]

AFFILIATIONS

[1]School of Mechanical and Electronic Engineering, Wuhan University of Technology, Wuhan 430070, China
[2]Hubei Provincial Engineering Technology Research Center for Magnetic Suspension, Wuhan 430070, China
[3]Laboratory for Future Interdisciplinary Research of Science and Technology, Tokyo Institute of Technology, 4259 Nagatsuta-cho, Midori-ku, Yokohama 226-8503, Japan

[a)]**Authors to whom correspondence should be addressed:** xguo@whut.edu.cn and knakamur@sonic.pi.titech.ac.jp.
[b)]**Present address:** Institute of Innovative Research, Tokyo Institute of Technology, Yokohama 226-8503, Japan.

ABSTRACT

Human activity recognition is widely used in many fields, such as the monitoring of smart homes, fire detecting and rescuing, hospital patient management, etc. Acoustic waves are an effective method for human activity recognition. In traditional ways, one or a few ultrasonic sensors are used to receive signals, which require many feature quantities of extraction from the received data to improve recognition accuracy. In this study, we propose an approach for human activity recognition based on a two-dimensional acoustic array and convolutional neural networks. A single feature quantity is utilized to characterize the sound of human activities and identify those activities. The results show that the total accuracy of the activities is 97.5% for time-domain data and 100% for frequency-domain data. The influence of the array size on recognition accuracy is discussed, and the accuracy of the proposed approach is compared with traditional recognition approaches such as k-nearest neighbor and support vector machines where it outperformed them.

Human activity recognition (HAR) has been widely used in practical applications, such as video surveillance, healthcare, wellbeing, smart homes, and sports. There are various ways of using different sensors for HAR, among which optical and acoustic methods are the conventional ways. In optics, HAR can be regarded as an automatic identification technology of human activity from images or videos. Methods based on RGB-D data were used for HAR[1,2] and proven to be outstanding approaches for human activity analysis. In recent years, deep learning algorithms have made the recognition of human gestures advance significantly, and the recognition rate was up to 99%.[3] However, the recognition methods of vision-based human activity are limited because of a lack of widespread acceptance in the private space, such as personal families due to privacy issues. Additionally, it is also difficult to be applied in special environments, such as dark, smoke, and dim light surroundings. These limitations can be avoided in acoustics. No matter in the special environment or in a small space, sound waves can propagate normally. In addition, acoustic methods to identify human activity will not cause privacy leaks and can be easily accepted by people. A single microphone or a few acoustic sensors were used in HAR. A microphone was used to recognize human behavior, and the mean value, median, autocorrelation, and other quantities were used to extract features of data. The recognition accuracy was not very high, and the average recognition rate was 89.6%.[4] A behavior recognition method based on eight sensors was studied, and it could quickly transform signal features into behavior feature vectors.[5] In the optical and acoustic sensing methods, various algorithms[6–8] are applied in HAR, convolution neural network (CNN), recurrent neural network (RNN), support vector machine (SVM), etc. They have a significant impact on recognition accuracy. In these algorithms, there were two kinds of information used for feature extraction. One was human geometry or motion information, which was used for establishing a model of human body features in order to improve the recognition rate.[9] The other was time-frequency information,[10] which was used for recognizing acoustic signals by deep learning methods. Although researchers have made outstanding contributions in the field of acoustics-based human behavior recognition, they often use a single microphone or several acoustic sensors to collect information that is normally one-dimensional. A few features are difficult to distinguish different types of activities for the one-dimensional information. In order to improve recognition accuracy, it is indispensable to increase features for a few sensors.

In this work, a two-dimensional acoustic array is used to collect human activity using low-level features extracted from the three-dimensional (3D) acoustic data. An approach based on a two-

dimensional acoustic array and CNN is proposed for HAR using a single feature extracted from the 3D acoustic data. The flowchart of the proposed approach is shown in Fig. 1. It basically consists of three parts, data acquisition and preprocessing, feature extraction, and activity recognition. The first part is data acquisition and preprocessing, four sensors as transmitters radiate ultrasonic waves with a frequency of 40 kHz, and 256 acoustic arrays (16×16) as receivers are used to collect reflected waves. Before the feature extraction, data preprocessing is performed to remove noise from adjacent channels interfering with each other. The second major part is the feature extraction. Standard deviation on the time- and frequency-sequential data is used to extract features. The final part is recognition that is performed to identify and classify types of human activity using deep learning methods.

The signals which carry information of human activity are collected by the acoustic array. Noise is generated in the collected signals, which is caused by electronic switching of different channels. Signals should be preprocessed to reduce noise before carrying out the next processing. The frequency of the transmitted signal is 40 kHz. A band-pass filter is used to allow signals with a frequency range of 35 to 45 kHz; thus, the impacts of noise can be reduced to a low level. Figure 2(a) shows the raw time-domain data of 16 sensors for standing activity, which includes noise. It is necessary to remove maximum values at the beginning for each sensor. After the preprocessing is implemented to all the raw datasets, the time-domain signals that contain valid information for each channel are obtained. The 2432 sampling points are obtained for 16 sensors in the time domain, which form one line of 3D data of human activity as shown in Fig. 2(b). In addition, Fourier transformation is used to obtain frequency-domain information.

In the time domain, the mean value of sampling points can be expressed as

$$m_t = \frac{1}{M}\sum_{i=1}^{M} x_i, \tag{1}$$

where m_t represents the mean value of sampling points, M is the number of sampling points, and x_i is the value of the ith sample. There is a

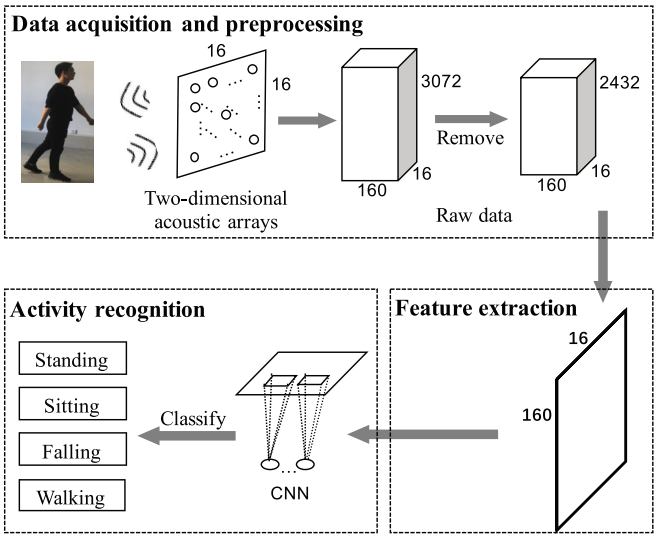

FIG. 1. Flowchart of the proposed approach.

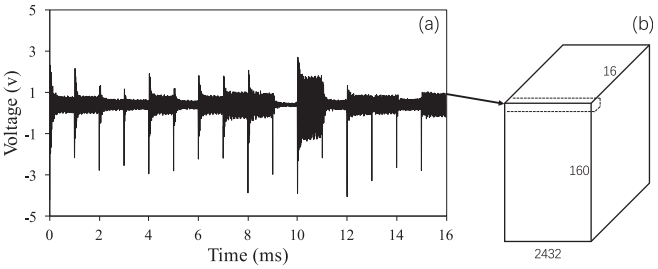

FIG. 2. (a) Raw time-domain data of 16 sensors for standing activity and (b) 3D data of human activity in the time domain.

significant fluctuation in the curve of two activities. Standard deviation of the signal can demonstrate the fluctuation of the sample point from the mean value. Therefore, standard deviation can be regarded as the eigenvalue of the signal and is given as

$$\sigma = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(x_i - m_t)^2}, \tag{2}$$

where σ is the standard deviation of M sampling points and represents the feature when analyzing the data characteristics in the time domain.

In the frequency domain, the difference of diverse activities on the amplitude-frequency diagram is analyzed, and the results show that the peak value of diverse activities is significantly different in the specific frequency domain around 40 kHz (39.3 to 40.7 kHz). The mean value m_f of its specific frequency amplitude is selected as the characteristic quantity and given as

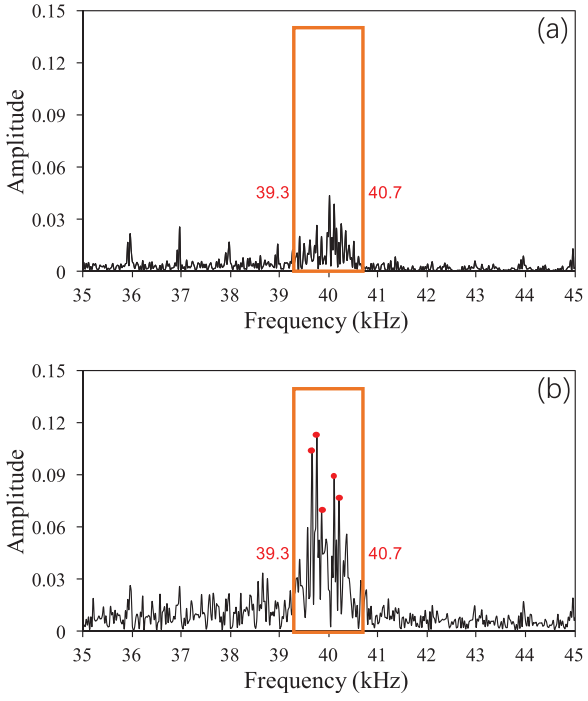

FIG. 3. Amplitude of no person (a) and standing activity (b).

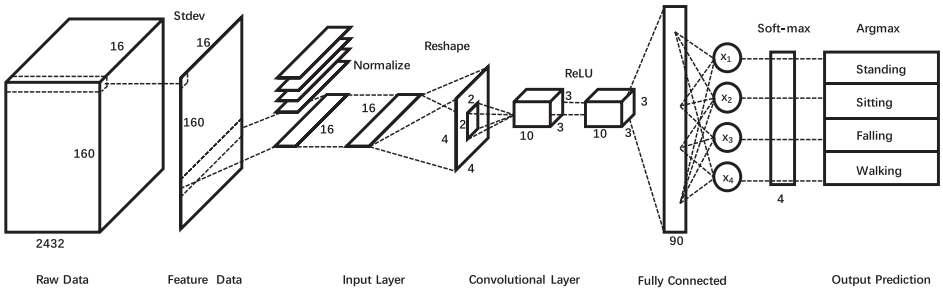

FIG. 4. Flowcharts of CNN processing for human activity recognition.

$$m_f = \frac{1}{N}\sum_{i=1}^{N} A_i, \tag{3}$$

where A_i is the values of amplitude. In this work, N represents the number of largest amplitudes in the range of 39.3 to 40.7 kHz and is equal to 5 according to experimental results. An example, amplitude of no person and standing activity for 5 peaks, is shown in Fig. 3.

In this work, CNN is used for HAR. It consists of an input layer, multiple convolution layers and pooling layers, a fully connected layer, and an output layer. The flowcharts of CNN processing for HAR are shown in Fig. 4. The size of the time- and frequency-domain data is $160 \times 16 \times N$ (N is the number of points). High dimensional data are not suitable as input of neural networks. The raw data are compressed to reduce the dimension. The data for each row are calculated by standard deviation and obtained as a value; hence, there are 16 values corresponding to 16 rows of the array. The feature data become the size of (16, 160). Each set of data contains 16 values for one activity, which can be represented by a matrix of (16, 1). The matrix is normalized as

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}, \tag{4}$$

where x_{min} is the minimum data of the input matrix, x_{max} is the maximum one, and x is the element of the input matrix. The normalized matrix is (16, 1), and then, it is reshaped into (4, 4) to meet the need of CNN, where the input matrix should be a squared matrix.

In the convolutional layer, the convolution is performed between the input layer and the convolution kernel to obtain the output feature. The output is often called the feature map; it can be expressed as[11]

$$x_j^l = f\left(\sum_{i \in M_j} x_i^{l-1} * k_{ij}^l + b_j^l\right), \tag{5}$$

where "$*$" denotes the convolution calculation, $f(\cdot)$ denotes the activation function, x_j^l and x_i^{l-1} denote the jth and ith feature maps of the l and $l-1$ convolutional layers, respectively, k_{ij}^l denotes the convolution kernel, M_j represents a selection of input maps, and b is the bias of each output map. In this processing, the convolution kernel has a size of (2, 2) and the stride is 1. The activation function is used to map the input of the neuron to the output. Since the Rectified Linear Unit (ReLU) is efficient in computing and can make the system stable, it is chosen as

$$f(x) = \begin{cases} 0, & x < 0 \\ x, & x \geq 0. \end{cases} \tag{6}$$

Finally, the function of softmax is used to perform numerical processing on the output of the classifier and process it into a probability value as

$$s_i = \frac{e^{V^i}}{\sum_j^C e^{V^i}}, \tag{7}$$

where V^i is the output of the prestage output unit of the classifier. i represents the category index, and the total number of categories is C. S_i represents the probability value.

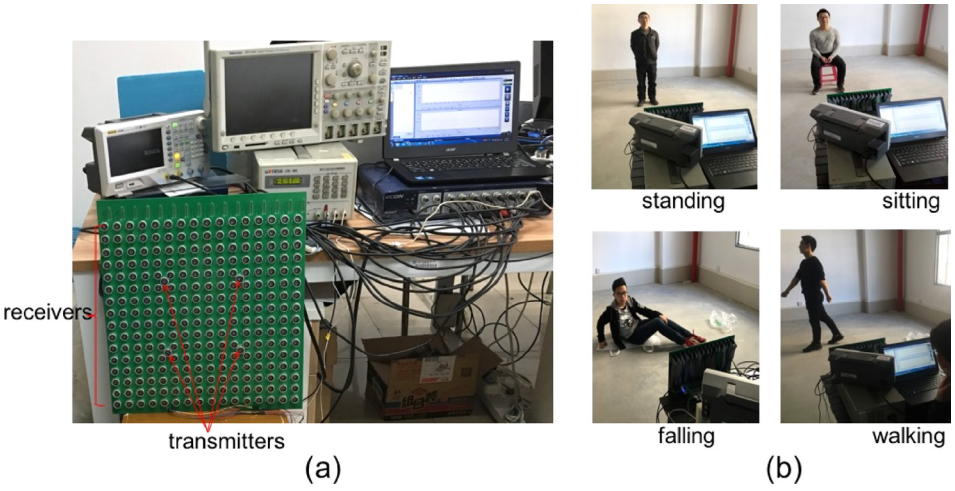

FIG. 5. (a) Photograph of the experimental setup and (b) four kinds of human activities.

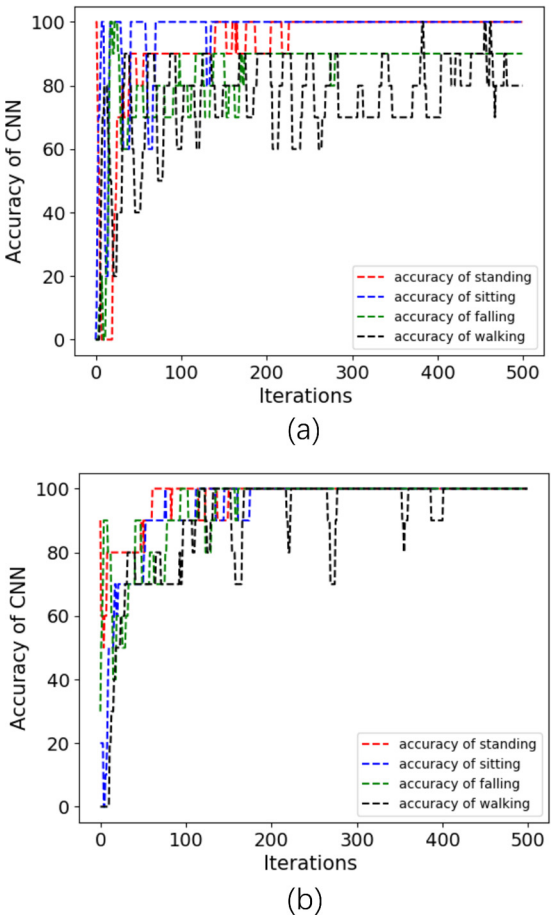

FIG. 6. (a) Accuracy of time-domain data by the CNN method and (b) accuracy of frequency-domain data.

Figure 5(a) shows the photograph of the experimental setup, which consists of 5 parts: a two-dimensional acoustic array, a digital oscilloscope, a function generator, a digital acquisition card, and a constant voltage source. The acoustic array has 256 (16×16) receivers and four ultrasonic transmitters in the center of board. The grid spacing of receivers is 25 mm, and the total dimensions of the array aperture are 395×395 mm. The array is used to transmit a high-frequency sinusoidal signal (40 kHz) and receive the reflected signal. For each

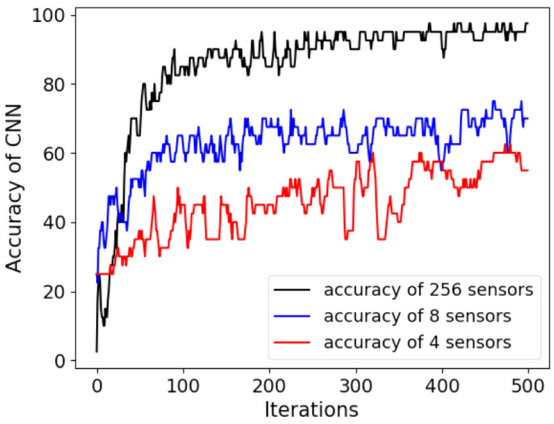

FIG. 7. Accuracy of three scale sensors.

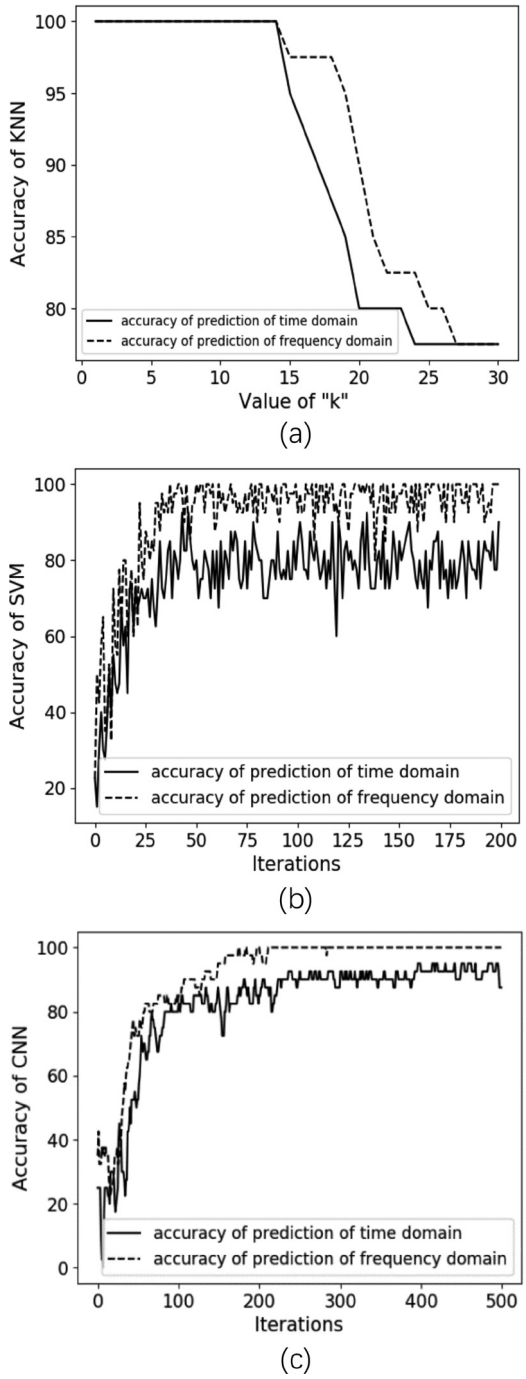

FIG. 8. Accuracy of KNN (a), SVM (b), and CNN (c) of time- and frequency-domain data.

period, the collection begins with the first row that collects data at the same time and then scans to the next row with an interval of 1 ms. It is finished until the data of 16 rows are collected. The experiment is carried out in a house. Effectiveness of the obtained data depends on characteristics of transmitters. The transmitters with an effective distance of 4 m and a directivity of 45° are used in experiments. Four volunteers stand at 2 m away in front of the array. They have a difference in their height and weight conducting four types of human activities: standing, sitting, falling, and walking, as shown in Fig. 5(b). Each individual performs each activity with 10 sets of data. For each activity, 25% of the data are used as a test set and the rest as a training set.

The test data are input into the CNN. The recognition rate of each activity is shown in Fig. 6. In the time domain as shown in Fig. 6(a), the accuracy of the activities (standing and sitting) is the highest in the four activities. This may be that the two activities are simple and in the static situation, and the activities (falling and walking) are in the moving situation. In the frequency domain as shown in Fig. 6(b), the accuracy of the activities (standing, sitting, and falling) is up to 100% less than 200 iterations and the accuracy of walking is also up to 100% around 400 iterations.

In the above results, high accuracy was obtained by the acoustic array with 256 sensors. A large number of sensors have a high requirement of hardware and more computation. Consequently, it is necessary to discuss the relationship between the number of sensors and recognition accuracy. In the time domain as shown in Fig. 7, the accuracy of three scale sensors is 62.5% for 4 sensors, 75% for 8 sensors, and 97.5% for 256 sensors, respectively. The results demonstrate that the scale of sensors can affect the accuracy, and the recognition accuracy of human activities can be significantly improved by increasing the number of sensors. The two-dimensional acoustic array has high accuracy compared with single or several sensors. This method can significantly reduce the complexity of feature extraction and also greatly reduce the signal processing workload and computational cost.

We also compare the recognition accuracy of KNN and SVM with CNN. From Fig. 8(a), it can be seen that the recognition accuracy of KNN can reach 100% in both the time domain and the frequency domain when the k value is less than 14. However, when the k value becomes higher, the accuracy of recognition drops dramatically. One possible reason is that the size of training data is relatively small. The KNN algorithm is just applied in small samples. From Fig. 8(b), it can be found that with the increasing number of iterations, the recognition accuracy increases rapidly and remains stable, and the frequency-domain recognition accuracy is 100%, which is significantly higher than that of 92.5% in the time domain. In Fig. 8(c), the total accuracy of the four activities is 97.5% for time-domain data. Recognition accuracy of frequency-domain data is higher than that of time-domain data and approaches 100%. The results show that accuracy in the frequency domain is higher than that in the time domain among the three algorithms. Meanwhile, it should be noted that the algorithms KNN and SVM are inferior to CNN in terms of overall recognition accuracy.

In this work, an approach has been proposed for HAR using a two-dimensional acoustic array and CNN. The standard deviation of data was used as a feature quantity. The extraction features were utilized as input vectors of CNN for activity training and recognition. The experiments have been carried out for four different types of human activities where an overall accuracy of 97.5% in the time domain and 100% in the frequency domain has been obtained. The influence of the different-scale acoustic arrays on the recognition rate was discussed. The results showed that the recognition rate of 256 sensors was 97.5%, which was much higher than that of 8 or 4 sensors. In addition, the accuracy of the proposed method was compared with that of traditional KNN and SVM approaches where it showed its superiority. In the future, we plan to focus on collecting more datasets in order to identify human activities more effectively in practical applications.

This work was supported by the National Natural Science Foundation of China (Grant No. 11504282).

REFERENCES

[1]P. Wang, W. Li, P. Ogunbona, J. Wan, and S. Escalera, Comput. Vision Image. Understanding **171**, 118 (2018).
[2]G. Zhu, L. Zhang, P. Shen, and J. Song, Sensors **16**, 161 (2016).
[3]J. C. Núñez, R. Cabido, J. J. Pantrigo, A. S. Montemayor, and J. F. Vélez, Pattern Recognit. **76**, 80 (2018).
[4]M. M. Hassan, M. Z. Uddin, A. Mohamed, and A. Almogren, Future Gener. Comput. Syst. **81**, 307 (2018).
[5]L. Köping, K. Shirahama, and M. Grzegorzek, Comput. Biol. Med. **95**, 248 (2018).
[6]A. Murad and J.-Y. Pyun, Sensors **17**, 2556 (2017).
[7]I. McLoughlin, H. Zhang, Z. Xie, Y. Song, and W. Xiao, IEEE-ACM Trans Audio Speech Lang. **23**, 540 (2015).
[8]G. Y. Liu, D. Y. Kong, S. G. Hu, Q. Yu, Z. Liu, T. P. Chen, Y. Yin, S. Hosaka, and Y. Liu, Appl. Phys. Lett. **113**, 084102 (2018).
[9]S. González, J. Sedano, J. R. Villar, E. Corchado, Á. Herrero, and B. Baruque, Neurocomputing **167**, 52 (2015).
[10]I. McLoughlin, H. Zhang, Z. Xie, Y. Song, W. Xiao, and H. Phan, PloS One **12**, e0182309 (2017).
[11]J. Bouvrie, Notes on convolutional neural networks (2006).