# Postprint: Applied Research on Convolutional Neural Networks for Quality Identification of Musical Instrument Plate Materials

**Authors:** Huang Yinglai, Li Xiaoshuang, Zhao Peng

**Date:** 2018-05-02T00:00:00+00:00

## Abstract

Current recognition algorithms for vibration signals of ethnic musical instrument panels suffer from disadvantages such as complex feature extraction and time-consuming processes. To address this issue, a wood vibration signal classification and recognition algorithm based on convolutional neural networks is proposed, enabling discrimination of panel quality for musical instruments. Convolutional neural networks combine the feature extraction and classification processes for neural network training, offering advantages such as high recognition accuracy and good robustness. First, the extraction of spectrogram features from wood vibration signals is analyzed and discussed in detail, followed by the application of convolutional neural networks combined with grid search methods for parameter optimization. To prevent overfitting, novel techniques such as ReLU and Dropout are also applied to obtain the final classification results. Experimental results demonstrate that the test sample accuracy reaches 96%, which is significantly superior to traditional methods. This method can reduce errors from manual measurement, accelerate panel selection time, and provides a more practical approach for material selection in the field of ethnic musical instrument manufacturing.

## Full Text

## Preamble

### Research on Wood Quality for Musical Instrument Recognition Using Convolutional Neural Network

*Huang Yinglai, Li Xiaoshuang, Zhao Peng*
(College of Information and Computer Engineering, Northeast Forestry University, Harbin 150040, China)

**Abstract:** Current vibration signal recognition algorithms for national musical instrument plates suffer from complex feature extraction and time-consuming processing. To address these issues, this paper proposes a wood vibration signal classification and recognition algorithm based on convolutional neural networks (CNN) to discriminate the quality of musical instrument plates. CNN combines feature extraction and classification processes for neural network training, offering advantages such as high recognition accuracy and robustness. The study first focuses on analyzing and discussing spectrogram features extracted from wood vibration signals, then applies CNN combined with grid search for parameter optimization. To prevent overfitting, novel techniques such as ReLU and Dropout are employed to obtain final classification results. Experimental results demonstrate that the test sample accuracy reaches 96%, significantly outperforming traditional methods. This approach can reduce manual measurement errors, accelerate plate selection time, and provide a more practical method for material selection in the national musical instrument manufacturing field.

**Keywords:** convolutional neural network; grid search; spectrogram; wood vibration signal

## 0 Introduction

In recent years, with China's rapid economic development and improved material living standards, people have developed a growing appreciation for national musical instruments. The beautiful melodies produced by these instruments not only cultivate character and broaden artistic horizons but also help alleviate high-intensity work stress. This has even sparked significant interest among foreigners, leading to a certain degree of development in China's national musical instrument industry. Chinese national musical instruments have a long developmental history and serve as carriers of traditional Chinese culture, possessing rich cultural heritage and ethnic characteristics, with increasingly diverse structures and varieties. As human requirements for various aspects of musical instruments continue to evolve, many national musical instruments rely heavily on wood, whose acoustic vibration properties substantially affect instrument quality. Therefore, research and application in identifying the quality of musical instrument plates hold important practical significance.

Currently, commonly used methods for identifying the quality of musical instrument plates both domestically and internationally include: Li Yuanzhe and Li Xianze from the Wood Industry Research Institute of the Chinese Academy of Forestry, together with Wang Xiquan and Wang Shuqin from the Beijing Musical Instrument Research Institute, proposed a method using acoustic wave excitation to study the acoustic properties of 31 major tree species in China; Dr. Shen Juan and Professor Liu Yixing from Northeast Forestry University conducted comprehensive and systematic research on the acoustic vibration

characteristics of spruce wood with different microfibril angles, analyzing the variation patterns of various acoustic parameters and their interrelationships, revealing how different structural factors affect the vibration performance of spruce wood; Ze Yuanjing evaluated wood acoustic properties by measuring parameters such as dynamic elastic modulus $E'$, flexural vibration internal friction $Q^{-1}$, and static bending elastic modulus $E$; Sobue, Tonosaki, and others used parameters including dynamic elastic modulus, radiation damping constant $R$, sound velocity, acoustic characteristic impedance $\omega$, dynamic loss tangent $\tan\delta/E$, and $\tan\delta$ to demonstrate that spruce is the optimal material for musical instrument soundboards; Treu et al. employed elastic testers and stress wave non-destructive evaluation techniques to grade spruce wood for instrument soundboards. Although these methods have achieved quality selection for musical instrument plates, most focus on experimental and analytical aspects of wood acoustic attribute parameters, with processes that are complex and time-consuming—these represent the limitations of traditional approaches.

With the development of deep learning, computer technology, and sound recognition technology, new methods can be considered to achieve faster and more accurate instrument material selection. Literature review reveals very few studies applying deep learning to wood vibration sound signal recognition, so this paper attempts to apply deep learning convolutional neural network methods to wood vibration signal classification and recognition.

Deep learning is a technology developed based on artificial neural networks. With the rapid development of deep learning, breakthrough progress has been achieved in the field of sound recognition. Convolutional Neural Networks (CNN) have become one of the popular research directions in image and sound recognition, possessing unique network structures that introduce "convolution" and "downsampling" operations to process multi-dimensional input features. CNN was actually proposed before 2006, but at that time it performed well only on small image recognition tasks and was not suitable for large-scale data. In 2012, Professor Hinton and his students achieved unexpectedly good results on the world-famous ImageNet competition using a deeper CNN model, marking CNN's gradual dominance in the field of image recognition. Around 2014, Dr. Dennis from Nanyang Technological University proposed spectrogram image features for sound, inspired by Zue's "spectrogram reading," achieving excellent recognition results and marking a breakthrough in sound event recognition. The University of Toronto and IBM Watson Research Group tested CNN models for sound input feature classification and found that using filter bank coefficients as sound features yielded the best recognition results. Hamid et al. combined NN/HMM with CNN for speech recognition, achieving success on the LVCSR database. Therefore, CNN-based image recognition and sound recognition technologies have received widespread attention and become current research hotspots.

The knocking sound of wood samples is a typical transient sound. Wood is an elastic solid material that can transmit sound wave energy through its elas-

tic medium. Under impact forces, wood can radiate acoustic energy through its own vibration, producing beautiful musical tones. The better its acoustic performance, the more excellent its acoustic resonance characteristics, which is an important basis for wood's widespread application in musical instrument manufacturing. The key to sound classification and recognition lies in feature extraction, and spectrogram features of sound signals have strong practical value. Spectrograms provide multiple characteristic information about wood vibration sounds, dynamically displaying changes in signal spectra and fully reflecting the spectral characteristics and complete information of sound signals. Therefore, this paper proposes extracting spectrogram features from wood vibration sound signals and applying convolutional neural networks for recognition, which is a new method that meets this requirement.

## 1 Convolutional Neural Network Structure

Convolutional Neural Network is a classic feedforward neural network that includes both forward propagation and backpropagation processes. CNN generally consists of convolutional layers, pooling layers, and fully connected layers. [Figure 1: see original paper] shows a classic CNN structure, LeNet-5.

### 1.1 Convolutional Layer

The convolutional layer, also called Conv.layer, forms the foundation of CNN. It comprises multiple feature maps and performs convolution operations directly on the original input signal (generally two-dimensional signals, images). A convolution kernel is a weight matrix whose size can be set arbitrarily, typically choosing 3×3 or 5×5 templates. The convolution kernel "slides" over the feature map with a certain stride, performing a convolution operation at each position. Through multiple convolution operations, different features of the input signal are extracted. Each convolution kernel can extract one type of feature, so $n$ convolution kernels can extract $n$ types of features. The typical calculation form of a convolutional layer is shown in Equation (1):

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

where $M$ represents the set of input feature maps, $k$ is the convolution kernel, $b$ is the bias term, $l$ represents the layer number, and $f(\cdot)$ represents the activation function (which can be nonlinear functions such as Sigmoid or Tanh).

### 1.2 Pooling Layer

The pooling layer is the input layer to the convolutional layer, also known as the downsampling layer of images. Pooling layers are typically inserted periodically between consecutive convolutional layers. Their function is to reduce computation and parameters in the network, thereby controlling overfitting. This paper

uses pooling functions to further adjust the output of this layer, where a pooling function replaces the network output at a certain position with the overall statistical characteristics of adjacent outputs. For example, the max pooling function used in this paper gives the maximum value within adjacent rectangular regions. Other commonly used pooling functions include average values within adjacent rectangular regions, L2 norm, and weighted average functions based on distance from the central pixel. The general form of a pooling layer is shown in Equation (2):

$$x_j^l = f\left(\beta_j^l \cdot p(x_j^{l-1}) + b_j^l\right)$$

where $\beta$ represents the weight coefficient and $p(\cdot)$ represents the pooling function.

### 1.3 Fully Connected Layer

Each neuron in the fully connected layer connects to all neurons in the previous layer. After multiple convolutional and pooling layers, one or more fully connected layers are connected to output the final classification results. The output values of the last layer are passed to the output layer. For the multi-classification problem in this paper, a softmax layer can obtain the probability distribution of the current sample belonging to different categories.

### 1.4 Other Common Layers

There are also some commonly used layers in CNN network structures, such as activation layers and Dropout layers. Generally, if the training dataset is not large, the training process may learn the noise in the training data, leading to overfitting. The specific manifestation is a large gap between model performance on the training set and test set, with weak model generalization ability. Therefore, adding these two layers in the network training part can effectively prevent overfitting.

This paper's activation layer uses the Rectified Linear Unit (ReLU) activation function, which "compresses" input values to the range of 0 to positive infinity. When the input value is positive, the ReLU function passes it directly; when the input value is negative, the ReLU function sets it to zero. In addition, there are Tanh and Sigmoid functions. In comparison, the ReLU activation function converges much faster, accelerating network training speed, as it only requires a threshold to obtain the activation value without computing large amounts of complex operations. The ReLU activation function is a piecewise function with the mathematical form shown in Equation (3):

$$ReLU(x) = \max(0, x)$$

Clearly, from the formula, when the input signal is less than 0, the output is always 0; when the input signal is greater than 0, the output equals the input. This shows that the ReLU function makes the output of some neurons 0, not only making the network sparse but also reducing dependencies between parameters, effectively alleviating overfitting.

The Dropout layer is typically placed after pooling layers and fully connected layers and is an excellent method to prevent overfitting. This paper adds Dropout technology in the network training part, which is equivalent to transforming the original network into a sparser network for continued training. Each node is randomly set to zero with a certain probability $p$. If $p$ is set to 0.5 in the experiment, it means randomly dropping 50% of neurons. Since network nodes do not respond to the immediate state of other nodes, this prevents certain nodes from functioning only under specific other nodes. A neural network with $n$ units can be seen as a collection of $2^n$ possible sparse neural networks. These networks share weights, so the total number of parameters remains $O(n^2)$ or fewer. For each presentation of each training model, a new sparse network is sampled and trained. Therefore, training a neural network using Dropout technology can be seen as training a collection of $2^n$ refined networks with extensive weight sharing. The network's generalization ability is improved, showing better adaptability.

## 2 Feature Extraction

Currently, commonly used sound feature parameters in sound recognition include Linear Predictive Cepstral Coefficients (LPCC) and Mel Frequency Cepstral Coefficients (MFCC). MFCC features are proposed based on human auditory characteristics, combining human hearing mechanisms with sound production mechanism characteristics. They do not rely on the assumption of an all-pole sound production model, correspond nonlinearly to the actual frequency of sound signals, simulate human ear sound processing characteristics to some extent, and demonstrate good robustness in noisy environments. In recent years, MFCC has been widely applied in various types of acoustics and achieved good results. LPCC is based on articulation models and does not fully consider human auditory characteristics, resulting in poor robustness in noisy environments. Although MFCC has obvious advantages over LPCC features, based on the algorithm considerations in this paper, spectrogram features of sound signals are more advantageous. Spectrograms can retain more information (including possible redundant information) and can use convolution and pooling operations to represent and process some typical sound invariances and variabilities. Therefore, this paper focuses on extracting spectrogram features.

A spectrogram is a sound spectrum diagram that converts audio into spectrogram images. This approach not only utilizes knowledge of speech signal processing but also integrates image processing technology, applying image processing techniques to speech processing. Therefore, applying spectrograms to sound event recognition is very promising. Spectrograms display a large amount of

sound-related feature information, with the horizontal axis representing time and the vertical axis representing frequency. The point corresponding to coordinates $(x, y)$ represents the speech data energy (i.e., sound intensity) at time $x$ and frequency $y$, with energy represented by color intensity—the darker the color, the stronger the speech energy at that point. This uses a two-dimensional plane to express three-dimensional information. Spectrograms have strong practical value, integrating time-domain waveform and spectrum features, clearly showing how sound features change over time, which cannot be displayed in waveform diagrams, as shown in [Figure 2: see original paper].

The entire extraction process of spectrograms is shown in [Figure 3: see original paper]. The sound signal contains noise and silent segments beyond wood vibration sounds, which affect sound feature performance. Noise can weaken some effective information in the signal, while silent segments affect the position of sound signals in the spectrogram. Therefore, to make experimental results more accurate, noise reduction and silence clipping are required for the sound. The original spectrogram and processed spectrogram are shown in [Figure 4: see original paper] and [Figure 5: see original paper].

This paper uses STFT to convert sound signals into spectrograms for feature extraction and crops the blank parts of the spectrogram to reduce computation and improve matching accuracy. First, pre-emphasis processing is applied to the sound signal to enhance the high-frequency part of the wood vibration sound, which is more conducive to spectral analysis of the entire vibration sound signal. The method uses a first-order high-pass filter with the mathematical expression:

$$H(z) = 1 - \mu z^{-1}$$

where $\mu$ is the pre-emphasis coefficient with a value close to 1, and $\mu = 0.97$ in this experiment.

Then, framing, windowing, and Short-Time Fourier Transform (STFT) processing are performed. The STFT is defined as:

$$X(n, \omega) = \sum_{m=0}^{M-1} x(m)\omega(n-m)e^{-j\omega m}, \quad 0 \le n \le N-1$$

where $n$ is the time-domain sampling point sequence, $n = 0, 1, \ldots, N-1$ ($N$ is the signal length); $x_m(n)$ is the sound signal after framing, $m = 0, 1, \ldots, M-1$, where $m$ is the frame synchronization time index and $n$ is the frame index; $\omega(n)$ is the Hamming window function that can reduce sound discontinuity caused by windowing operations, defined as:

$$\omega(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{L-1}\right) & 0 \le n \le L-1 \\ 0 & \text{other } n \end{cases}$$

where $L$ is the window length. Generally, when the frame length is 10-30ms, the signal can be considered stationary. This experiment uses a frame length of 512 points, with the overlap between frames being the frame shift, generally half the frame length, i.e., 256 points.

Next, the logarithmic energy method is used to generate the spectrogram:

$$\hat{X}(x, y) = \log(|S(x, y)|)$$

where $S(x, y)$ represents the time-frequency matrix.

Finally, the mapminmax method is used to normalize the spectrogram. Normalization standardizes the data, making its grayscale variation range $[0, 1]$, thereby ensuring uniform statistical distribution of sample data for more accurate and convenient subsequent processing. The normalization process is defined as:

$$G(x, y) = \frac{S(x, y) - \min(S(x, y))}{\max(S(x, y)) - \min(S(x, y))}$$

where $\min(S(x, y))$ and $\max(S(x, y))$ represent the minimum and maximum values in the time-frequency matrix, respectively. After normalization, the time-frequency matrix $G(x, y)$ within $[0, 1]$ is obtained and used as grayscale intensity values to generate the grayscale spectrogram.

Extracting features from wood vibration signals is a critical step for further analysis and recognition of wood quality, directly affecting the recognition speed and accuracy of subsequent experiments.

## 3 Experiments and Results

### 3.1 Experimental Sound Sample Set and Environment

The experimental sound data in this paper were obtained by striking Paulownia wood plates of the same size but different quality grades in a quiet laboratory. The collection device was a voice recorder. Repeated acoustic acquisition was performed at different positions on three different quality grades of wood plates to obtain single-channel 16-bit wav format sound files with a sampling rate of 44.1KHz. CoolEdit software was then used to cut and filter the original sound samples to obtain individual sound samples for subsequent experiments. The test sample format was 116×80 pixel single-channel grayscale images, with 7,319 training samples and 2,196 test samples. Detailed information about the experimental data is shown in .

The latest neural network library Keras has attracted widespread attention, using Theano or TensorFlow as backends. This paper uses MATLAB 2016a programming software for feature extraction from original wood vibration signals, and the built CNN runs on a Linux platform with Keras on a TensorFlow backend for Python 3.5.

### 3.2 Experiments and Results Analysis

**3.2.1 CNN Model Adjustment Based on Grid Search** Since the dataset in this paper is not large, a grid search-based method is considered to adjust different parameter settings to fit the CNN model, thereby avoiding arbitrary and blind parameter selection. This paper optimizes parameters including optimizer, Dropout rate, number of epochs, and batch_size through grid search. Experimental parameter settings and results are shown in , leading to the following conclusions:

a) Batch size that is too small (e.g., 1) hinders network convergence, while batch size that is too large reduces the number of iterations, thereby requiring longer time to achieve good accuracy. Based on program results, the batch size is recommended to be set to 64.

b) Regularization: Dropout is a simple and commonly used regularization technique suitable for preventing overfitting. The Dropout rate was tested from 0.5 to 0.2, and a value of 0.3 is recommended based on testing.

c) Number of epochs refers to the number of times the training set is input into the neural network for training. When the difference between test error rate and training error rate is small, the current number of epochs is considered optimal; otherwise, the network structure needs adjustment or the number of epochs increased. This paper selects 10 epochs as the optimal value.

d) There are many types of optimizers, such as SGD, RMSprop, Adadelta, and Adam. After experimental testing, the Adam optimizer is selected for subsequent experiments.

**3.2.2 Comparison Experiments of Different CNN Network Structures** The CNN built in this paper adopts three types of structures with the following parameter settings (convolutional layers denoted by C, pooling layers by P, and fully connected layers by F). Activation layers (ReLU) are added after each convolutional layer, and Dropout layers (Dropout rate of 0.3) are added after each pooling layer:

a) "C1+P1+F1+F2" structure: C1 layer has 16 convolutional kernels of size 3×3; P1 size is 2×2 using max pooling output; F1 layer has 64 nodes, and F2 layer has 3 nodes (output categories), denoted as CNN-1.

b) "C1+P1+C2+C3+P2+F1+F2" structure: Other layers are the same as CNN-1. C2 layer has 32 convolutional kernels of size 3×3; C3 layer has 32 convolutional kernels of size 5×5; P2 size is 2×2 using max pooling output, denoted as CNN-2.

c) "C1+C2+P1+C3+C4+P2+F1+F2" structure: Other layers are the same as CNN-1. C2 layer has 16 convolutional kernels of size 3×3; C3 layer has

32 convolutional kernels of size 3×3; C4 layer has 32 convolutional kernels of size 5×5; P2 size is 2×2 using max pooling output, denoted as CNN-3.

The three convolutional structures were trained and tested for classification effectiveness, with results shown in [Figure 6: see original paper]–[Figure 11: see original paper]. Based on these results, CNN-2 achieves higher classification recognition rates with the smoothest accuracy and loss curves and moderate experimental running time. CNN-1, despite having the shortest experimental time and high test set recognition rate, has low training set recognition rate. CNN-3 shows some irregular curves, lower recognition rates compared to CNN-1 and CNN-2, and the longest running time. One epoch represents a complete iteration (all samples are trained), and this paper uses 10 iterations. In the figures, loss represents training set loss value, acc represents training set accuracy; val_loss represents test set loss value, and val_acc represents test set accuracy. In the CNN-2 model, the last iteration can converge to 96% prediction accuracy. Detailed accuracy, recall, and other parameters for each wood vibration sound category are shown in .

Comprehensive comparison shows that CNN-2 achieves the best experimental results. Structures with fewer layers require less running time but come with more weights, which increases memory burden—this is why CNN-2 outperforms CNN-1. A single convolutional kernel can only detect one feature (e.g., image texture, vertical edges), while features in spectrograms are often more complex, making one convolutional kernel insufficient. Therefore, convolutional layers in neural networks have multiple convolutional kernels, with different feature maps containing different features, resulting in multi-layer convolutional layer outputs. The number of convolutional kernels is generally set to increase by even multiples. More feature maps extract more features, but more convolutional kernels require processing more parameters. To reduce model complexity, the number of convolutional kernels in this paper does not exceed 32. The CNN-3 network structure is more complex than CNN-2, with one additional convolutional layer containing 32 kernels, requiring more parameters and lower training efficiency, without improved accuracy. This demonstrates that simply increasing the number of convolutional kernels and network complexity does not correspondingly improve classification performance. Other factors such as good network empirical parameters, appropriate number of iterations, and batch values also affect classifier performance. Spectrograms can effectively represent features of vibration sound signals from wood of different quality grades. Compared with other networks, CNN's downsampling operation has scale translation invariance, reducing feature map dimensions and avoiding overfitting. CNN's local region perception operation can reduce noise, enhance speech features, and better analyze energy distribution.

## 4 Conclusion

This paper introduces deep learning convolutional neural network methods for wood vibration signal recognition and proposes extracting spectrogram features

from wood vibration signals combined with image recognition technology. The study also investigates parameter optimization and CNN structure, applying CNN as a classifier in the recognition system. Grid search technology is used to obtain optimal parameters. Experimental results demonstrate that the proposed algorithm has good effectiveness and is more scientific and practical compared with previous detection methods. The experiment still has limitations in sample quantity and network performance. In future research, more sound vibration signal data samples of Paulownia wood will be collected, and network structures will be further improved to enhance recognition accuracy.

## References

[1] Luo Jiaqi, Tang Heng. Inheritance and development of Chinese national musical instruments [J]. Forward Position, 2013, 30(7): 166-168.
[2] Li Yuanzhe, Li Xianze, Wang Xiquan, et al. Measurement of acoustic properties of several wood species [J]. Forestry Science, 1962, 7(1): 59-66.
[3] Shen Juan, Liu Yixing, Liu Zhenbo, et al. Effect of microfibril angle on acoustic vibration characteristics of spruce wood [J]. Journal of Northeast Forestry University, 2002, 30(5): 50-52.
[4] Ono T, Norimoto M. Study on Young's modulus and internal friction of wood in relation to the evaluation of wood for musical instruments [J]. Limnology & Oceanography, 1983, 22(4): 611-614.
[5] Tonosaki M, Okano T, Asano I. Vibrational properties of Sitka Spruce with longitudinal vibration and flexural vibration [J]. Mokuzai Gakkaishi, 1983.
[6] Sobue N. Measurement of Young's modulus by the transient longitudinal vibration of wooden beams using a fast Fourier transformation spectrum analyzer [J]. Journal of the Japan Wood Research Society, 1986, 32(9): 744-747.
[7] Treu A, Hapla F. Study on the quality of spruce and fir resonance wood [J]. Allgemeine Forst-und Jagdzeitung, 2000, 171(12): 215-222.
[8] Lecun Y, Bengio Y, Hinton G. Deep learning [J]. Nature, 2015, 521(7553): 436-444.
[9] Zhou Feiyan, Jin Linpeng, Dong Jun. Review of convolutional neural network research [J]. Chinese Journal of Computers, 2017, 40(06): 1229-1251.
[10] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural network [J]. Science, 2006, 313(5786): 504-507.
[11] Dennis J W. Sound event recognition in unstructured environments using spectrogram image processing [D]. Singapore: Nanyang Technological University, 2014.
[12] Dennis J, Tran H D. Enhanced local feature approach for overlapping sound event recognition [C]// Proc of Summit and Conference on Asia-Pacific Signal and Information Processing Association. 2015: 1-4.
[13] Zue V. Notes on speech spectrogram reading [C]// MIT Lecture Notes. Cambridge, MA, 1991.
[14] Sainath T N, Mohamed A R, Kingsbury B, et al. Deep convolutional neural networks for LVCSR [C]// Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. 2013: 8614-8618.

[15] Abdel-Hamid O, Mohamed A R, Jiang H, et al. Convolutional neural networks for speech recognition [J]. IEEE//ACM Trans on Audio Speech & Language Processing, 2014, 22(10): 1533-1545.
[16] Deng Liu, Wang Zijie. Research on vehicle type recognition based on deep convolutional neural network [J]. Application Research of Computers, 2016, 33(03): 930-932.
[17] Li Jian. Wood Science [M]. 3rd ed. Beijing: Science Press, 2014.
[18] Liang Shili, Wei Ying, Pan Di, et al. Speaker-dependent two-character Chinese vocabulary recognition based on spectrogram row projection [J]. Journal of Jilin University: Engineering and Technology Edition, 2017, 47(1): 294-300.
[19] Tao Huawei, Zha Cheng, Liang Ruiyu, et al. Spectrogram feature extraction algorithm for speech emotion recognition [J]. Journal of Southeast University: Natural Science Edition, 2015, 45(05): 817-821.
[20] LeCun Y. LeNet-5, convolutional neural networks [EB/OL]. 2015. http://yann.lecun.com/exdb/lenet.
[21] Srivastava N, Hinton G E, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting [J]. Journal of Machine Learning Research, 2014, 15(1): 1929-1958.
[22] Hinton G E, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing co-adaptation of feature detectors [J]. Computer Science, 2012, 3(4): págs. 212-223.
[23] Francois C. https://keras.io/ [EB/OL].
[24] Li Han, Yang Xiaofeng, Deng Hongxia, et al. Adaptive PCNN model parameters based on grid search algorithm [J]. Computer Engineering and Design, 2017, 38(1): 192-197.

*Note: Figure translations are in progress. See original paper for figures.*