Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 3686643, 14 pages
https://doi.org/10.1155/2022/3686643

Research Article

Application of Artificial Intelligence Elements and Multimedia Technology in the Optimization and Innovation of Teaching Mode of Animation Sound Production

Aijun He

School of Art and Design, Lanzhou Jiaotong University, Lanzhou, China

Correspondence should be addressed to Aijun He; heaijun@mail.lzjtu.cn

Received 18 January 2022; Revised 7 February 2022; Accepted 9 February 2022; Published 27 February 2022

Academic Editor: Kalidoss Rajakani

Nowadays, with the rapid development of multimedia technology and computer information processing, the data of multimedia information presents explosive growth. At present, the method of using artificial recognition of sound materials is inefficient, and an automatic recognition and classification system of sound materials is needed. To improve the accuracy of sound recognition, two algorithm models are established to identify and compare the sound materials, which are the hidden Markov model (HMM) and back propagation neural network (BPNN) model. Firstly, HMM is established, and the sound material is randomly selected as the test sample. The comparison between the expected classification and the actual is tested, and the recognition rate of each classification is got. The final average recognition rate is 61%. The anti-interference characteristics of the training HMM are tested, and the identification rate of the training model is selected in 6 types of signal-to-noise ratio (SNR) environments. The recognition rate of the training model has an obvious downward trend with the decrease of the SNR. Secondly, the BPNN model is built, and 200 BPNN training experiments are conducted. The training model with the highest average recognition rate is selected as the experimental model. The average recognition rate of the final model is higher than 90%. The expression ability and stability of the trained model are simulated after the new sample is introduced, and the anti-interference performance of the model is tested in different environments of SNR. The results of performance test are good, and only the recognition rate of complex types of some sound sources decreased. Finally, the accuracy of the HMM in the experiment is not as high as that obtained by BPNN. Therefore, the training method of BPNN has a greater advantage in both recognition accuracy and recognition efficiency for the studied sound. It provides a reference for automatic recognition of sound materials.

1. Introduction

In the production of animation sound, the dubbing part requires a lot of sound materials. Although a large number of materials are stored in the sound effect material library, some dubbings need to be temporarily designed to dub for it or to record by onomatopoeia through various props [1]. At present, audio creation is still inseparable from manual recognition, and sound effects contain a variety of available sound resources. Through editing, classification, mixing and other operations, its materials will be arranged very compactly and rely on human ears for hearing. There is no problem in the short term. If the amount of material is large, it will consume a lot of energy and lead to hearing fatigue. In serious cases, it will also cause memory errors, judgment deviations, etc. [2–4]. In the field of sound classification and recognition, speech recognition has developed relatively maturely, including listening to songs and recognizing music and other functions, which have been widely used. Nowadays, there are few studies related to the automatic classification of animated sound effects.

Haeb-Umbach et al. introduced the algorithm used to achieve accurate long-distance speech recognition. Although deep learning (DL) occupies a large share of technological breakthroughs, the ingenious combination with traditional signal processing can bring effective solutions [5]. Yu et al. proposed a peak-based framework for the errored second ratio (ESR) task from a perspective of more brain-like.

Results show that compared with other baseline methods, the experimental design framework performs the best. The peak-based framework has several favorable features, including early decision-making, small dataset acquisition, and continuous dynamic processing [6]. Jin et al. successfully prepared a sound detector based on MXene by combining DL with 2DMXenes, which had improved recognition and sensitive response to pressure and vibration, which helped to produce high recognition and resolution [7]. Li et al. studied the classification of feeding behavior of dairy cow based on automatic sound recognition and found that DL technology can classify feeding behavior [8]. Lhoest et al. indicated a classifier based on classical machine learning (ML) and a lighter convolutional neural network (CNN) model for environmental sound recognition. The results show that the classic ML classifier can be combined to obtain results similar to DL models and even better than DL models in accuracy [9]. Demir et al. explored the classification of environmental sound based on depth features. The depth feature is extracted by using the fully connected layer of a newly developed CNN model, which is trained with spectrogram images in an end-to-end manner. Experiments show that the classification accuracy of the model reaches 96.23% and 86.70%, respectively [10]. Zhang et al. studied the application of CNN and recurrent neural network units based on feature fusion in environmental sound classification and found that the model with load manage control center (LMCC) as input is suitable for solving problems of electronic stability control. The model can achieve good classification accuracy [11]. Catanghal advanced and discussed a framework of a detection system in the study room. Feature extraction technology is used to obtain the representation of parameter type, which is used to analyze the sound of the intelligent home machine listening system specially used in the study room. It is concluded that ML is feasible for sound detection and can be applied as a technology in an innovative learning environment [12–14].

By consulting the references, it shows that the current research is basically in the stage of feasibility analysis or the effect of classification and recognition is not obvious enough, and the efficiency and accuracy of automatic classification and recognition of sounds need to be improved. For this reason, the idea of applying artificial intelligence (AI) components and multimedia technology to sound recognition is proposed, which can realize the automatic recognition of sound materials and avoid the problems of labor time and inefficiency. Firstly, the sound feature recognition combined with ML is mainly aimed at fitting problems caused by different ML algorithms or improving small samples in the experiment. Secondly, the model used in the simulation is adjusted to the best performance through combination to achieve the effect of classification and recognition. High-efficiency sound recognition is designed through ML algorithm to facilitate the classification of sound materials.

2. Establishment of Algorithm Model

2.1. Algorithm Model Based on Hidden Markov Model. The hidden Markov model (HMM) is composed of the hidden Markov chain. HHM describes the process of state transition. For the first order of the HHM, state transition depends on several states in the system. The probability of state transition refers to the probability of one state to another. The probabilities of all transitions are represented through the matrix of state transitions, and this matrix will not change over time. The initial probability is the probability parameter of any state in the initial state of the model [15, 16]. Generally, HMM includes the matrix of the initial state and of state transition.

HMM is usually represented by $\lambda = (P, Q, V)$, and the completed HMM should also have two other parameters, N is the specified state parameter and M is the observation symbol. These two parameters and three density probabilities (P, Q, V) constitute the HMM.

G is the number of states in HMM and a collection of hidden states. When $s = \{S_1, S_2, \cdots, S_N\}$ in the collection of model states, and the state of t moment is represented as q_t, $q_t \in S_1, S_2, \cdots, S_N$.

H is the number of observations. The set of observation symbols is $O = \{O_1, O_2, \cdots, O_N\}$ in the model.

P is the probability distribution of state transition, which is a vector matrix composed of hidden transition probability. The state transition probability of the hidden Markov chain represents the probability of transition from one hidden state to another.

$$A = \{a_{ij}\}, a_{ij} = P[q_{t+1} = S_j \mid q_t = S_i], 1 \le i, j \le N. \quad (1)$$

In equation (1), a_{ij} has the following characteristics:

$$a_{ij} \ge 0, \sum_{j=1}^{N} a_{ij} = 1. \quad (2)$$

Q is the probability distribution of observation symbol in state S_j. A specific hidden state will generate a specific probability matrix of the observation state in the specified HHM. Therefore, in state S_j, the probability distribution of observation symbols includes the observation probability matrix obtained by specifying the hidden states, which can also be defined as a confusion matrix.

$$B = \{b_j(O_k)\}, b_j(O_k) = p, 1 \le j \le N, 1 \le k \le M. \quad (3)$$

V is the probability distribution in the initial state, as shown in

$$\pi = \{\pi_i\}, \pi = p[q = S_i], 1 \le i \le N. \quad (4)$$

These five parameters are generally called the five elements of the HMM, as shown in Figure 1.

Therefore, the HMM is to add the concept of observing state distribution in the conventional Markov process, and the probability distribution relationship between hidden states and observation states is also established in the actual algorithm of this model [17].

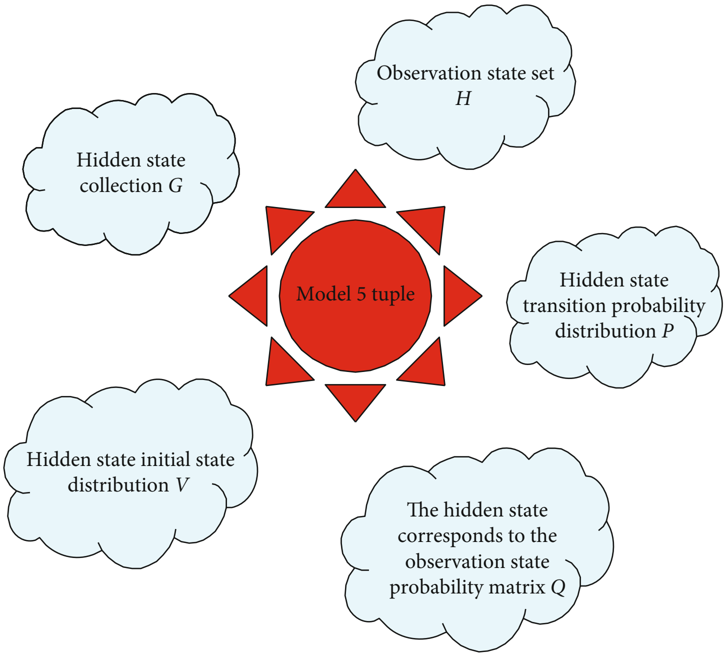

Observation state set H
Hidden state collection G
Model 5 tuple
Hidden state transition probability distribution P
Hidden state initial state distribution V
The hidden state corresponds to the observation state probability matrix Q

Figure 1: Five elements of the HMM.

If the HMM λ is given and the observation sequence of each part of t time is $O_1, O_2, \cdots, O_t$ and the state q_t is a forward probability, as shown in

$$a_t(i) = p(o_1, o_2, \cdots o_t, i_t = q_i \mid \lambda). \tag{5}$$

Forward probability $a_t(i)$ and observation sequence probability $p(0 \mid \lambda)$ can be obtained by recursive. The process is shown in

$$a_{t+1}(i) = \left[\sum_{j=1}^{N} a_t(j) a_{ji}\right] b_i(O_t + 1). \tag{6}$$

In equation (7), a_{ji} is the transition probability.

$$a_{ji} = p\left(i_{t+1} = q_i \mid i_t = q_j\right). \tag{7}$$

Combined with forward probability, the definition is shown in

$$a_t(j) = p\left(o_1, o_2, \cdots o_t, i_t = q_j \mid \lambda\right). \tag{8}$$

Combined with HMM, the hypothesis is shown in

$$a_{ji} = p\left(i_{t+1} = q_i \mid i_t = q_j\right) - p\left(i_{t+1} = q_i \mid o_1, o_2, \cdots o_t, i_t = q_j\right). \tag{9}$$

In

$$p\left(o_1, o_2, \cdots o_t, i_t = q_j, i_{t+1} = q_i \mid \lambda\right). \tag{10}$$

Through the summation processing, the equation is shown in

$$\sum_{j=1}^{N} a_t(j) a_{ji} = p\left(o_1, o_2, \cdots o_t, i_t = q_j, i_{t+1} = q_i \mid \lambda\right). \tag{11}$$

The observation probability in the recursive equation b_i (o_{t+1}), combined with the independent hypothesis of observation, is as

$$b_i(o_{t+1}) = P(o_{t+1} \mid i_{t+1} = q_i) = P(o_{t+1} \mid o_1, o_2, \cdots o_t, i_{t+1} = q_i). \tag{12}$$

$a_{t+1}(i)$ can be expressed in probability, as

$$a_{t+1}(i) = p(o_1, o_2, \cdots o_t, o_{t+1}, i_{t+1} = q_i \mid \lambda). \tag{13}$$

To get the value of $P(0 \mid \lambda)$, all forward probability of the last state of Markov sequence is summed, as

$$P(0 \mid \lambda) = \sum_{i=1}^{N} a_T(i) = P(o_1, o_2, \cdots o_t, i_{t+1} \cdots, o_T \mid \lambda). \tag{14}$$

If the HMM λ is given, under the condition of the state q_t at t time, the observation sequence of part from t to T is $O_{t+1}, O_{t+2}, \cdots, O_T$ and the state q_t is a backward probability, as shown in

$$\beta_t(i) = P(O_{t+1}, O_{t+2}, \cdots, O_T \vee i_t = q_i, \lambda). \tag{15}$$

According to the successive approximation algorithm, the Q function needs to be represented first. Through the parameters and observation variable conditions in the model, the logarithmic function of the data is relative to the hidden probability of variable condition, and the distribution expectation is Q function, as shown in

$$Q\left(\lambda, \widehat{\lambda}\right) = \sum_{i=1}^{N} \log P(0, I \mid \lambda) P\left(I \mid 0, \widehat{\lambda}\right). \tag{16}$$

In the equation, $\widehat{\lambda}$ is the current estimate of the HMM, which is the parameters of the maximum HMM. According to the successive approximation algorithm, and then the maximization, so the Q function needs to be decomposed and calculated.

In

$$\log P(0, I \mid \lambda) = \pi_{i_1} b_{i_1}(o_1) a_{i_1 i_2} b_{i_2}(O_2) \cdots a_{i_{T-1}} i_T b_{i_T} O_T. \tag{17}$$

In

$$P\left(I \mid 0, \widehat{\lambda}\right) = P\left(0, I \mid \widehat{\lambda}\right) \div P\left(0 \mid \widehat{\lambda}\right). \tag{18}$$

The estimated parameters appear in the three terms when they are substituted into the Q function, respectively, and only need to be maximized for each item.

$$\begin{aligned} Q\left(\lambda, \widehat{\lambda}\right) = & \sum_{i=1,j=1}^{N} \sum_{t=1}^{T-1} \log a_{ij} P\left(O, i_t = i, i_{t+1} = j|, \widehat{\lambda}\right) \\ & + \sum_{j=1}^{N} \sum_{i=1}^{T} \log b_j(O_t) P\left(O, i_t = j|, \widehat{\lambda}\right) \\ & + \sum_{i=1}^{N} \log \pi_i P\left(O, i_1 = i \mid \widehat{\lambda}\right). \end{aligned} \tag{19}$$

Transition from any hidden state to hidden state S_i means that for the sum of time t, including all time expectations in the grid, it corresponds to the expectations of the state i under observation O.

The essence of the Viterbi algorithm is to specify the observation sequence to find the maximum possibility of the state sequence, which is actually to maximize $P(I \mid 0, \lambda)$. First input δ and φ.

$$\delta_t(i) = \max_{i_1, i_2 \cdots i_t} P(i_t = i, i_{t-1}, \cdots, i_1, a_t, \cdots, a_1 \mid \lambda). \tag{20}$$

δ is the maximum probability in all single paths $(i_1, i_2, \cdots, i_t)$ with a state of i at t moment, and the variable δ is recursive, as shown in

$$\begin{aligned} \delta_{t+1}(i) & = \max_{i_1, i_2 \cdots i_t} P(i_{t+1} = i, i_t, \cdots, i_1, a_{t+1}, \cdots, a_1 \mid \lambda) \\ & = \max_{1 \le j \le N} \left[\delta_t(j) a_{ji}\right] b_i(O_{t+1}). \end{aligned} \tag{21}$$

Specify the starting value $\delta_1(i) = \pi_i b_i(O_1)$ and then iterate, and the termination result is shown in

$$p^* = \max_{1 \le j \le N} \delta_T(i). \tag{22}$$

The Viterbi algorithm stores a reverse pointer for any state. The partial probability will reach the specified state according to the reverse pointer. The calculation of partial probability in the Viterbi algorithm is different from that processed in the forward algorithm, because the probability will not change over time. The Viterbi algorithm calculates the probability of the most direct path of reaching a certain state at t moment, not the sum of all paths. When $t = 1$, there is no way to find the maximum possible path to reach a certain state. Then the initial probability of the state in which $t = 1$ is multiplied by the observation probability in the corresponding observation state to calculate the partial probability, which is similar to the forward algorithm. The result of partial probability is obtained by multiplying the initial probability and the observation probability [18].

The structure of the HHM consists of two closely related steps. One is an observable Markov chain, and the other is a hidden process that matches number of states and observations of the model. The states of HMMs can be transferred to each other over time, and they can also remain in one state. The training process uses audio clips of five seconds for each category, and the training flow chart is shown in Figure 2.

When processing audio of each category, the HMM consists of two closely connected processes, one is an observable Markov chain and the other is a hidden process that matches number of states and observations of the model. The states of HMMs can be transferred to each other over time, and they can also remain in one state. In this simulation, HHMs are established for each classification.

Using the Viterbi algorithm in logarithmic form, the initial and transition probabilities are calculated separately. The code is shown in this.

$$\begin{gathered} \log\,(\text{init}) \\ \text{ind1} = \text{find}(\text{init} > 0)\,; \\ \text{ind0} = \text{find}(\text{init}< = 0)\,; \\ \text{init}(\text{ind0}) = -\,\text{inf}\,; \\ \text{init}(\text{ind1}) = \log\,(\text{init}(\text{ind1}))\,; \\ \log\,(\text{trans}) \\ \text{ind1} = \text{find}(\text{trans} > 0)\,; \\ \text{ind0} = \text{find}(\text{trans}< = 0)\,; \\ \text{trans}(\text{ind0}) = -\,\text{inf}\,; \\ \text{trans}(\text{ind1}) = 1\,\log\,(\text{trans}(\text{ind1})). \end{gathered} \tag{23}$$

2.2. Establishment of Model Based on BPNN Algorithm. An artificial neural network (ANN) is an adaptive neural network composed of simple neurons. It has its own nonlinear characteristics and can simulate the human nervous system connected in parallel to perform qualitative and quantitative

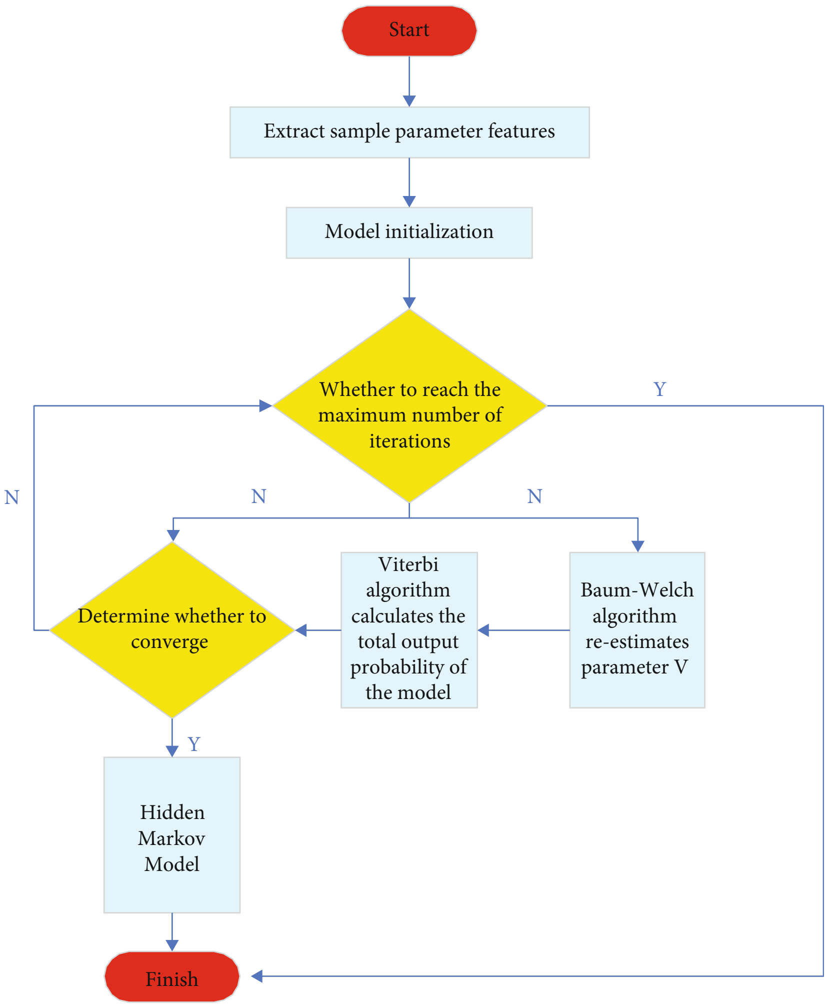

Start
Extract sample parameter features
Model initialization
Whether to reach the maximum number of iterations
Y
N
N
N
Determine whether to converge
Viterbi algorithm calculates the total output probability of the model
Baum-Welch algorithm re-estimates parameter V
Y
Hidden Markov Model
Finish

Figure 2: Training flowchart of HMM.

operations. Because it is actually composed of many neurons, in ANN, the output of one neuron is the input of another neuron. Forward propagation means that the signal passes through the input layer and through the operation of neurons to output. There are many hidden layers and output neurons in the neural networks (NNs). They are evolved through biological neuron models. In biological NNs, neurons will transmit chemicals to other neurons after feeling "excited." Neurons are linked to each other, and the rest of neurons will transmit information through incoming and out of nerves and finally handed over to the central nervous system processing to form NNs in machine learning [19–21].

Workflow of the output-perceived neuron receives input signal x_i at the input end. According to the link weight w_i, s is regarded as an external input signal. All input weights are shown in

$$\sigma = \sum_{i=1}^{n} w_i x_i + s. \tag{24}$$

Function f is a nonlinear feature function. Use this function to convert and get the output y:

$$y = f(\sigma). \tag{25}$$

In the equation, the f function is an activation function, and the reverse propagation neural network and deep learning (DL) usually use the S-type logarithmic function or tangent function. The expression of logarithmic S-type activation function is shown in (b is a deviation value)

$$f(x) = \frac{1}{1 + \exp\left(-(x+b)\right)}. \tag{26}$$

The expression of hyperbolic tangent S-type activation function is shown in (b is the deviation value)

$$f(x) = \frac{1 - \exp\left(-2(x+b)\right)}{1 + \exp\left(-2(x+b)\right)}. \tag{27}$$

The topology structure is formed by the links between neurons. The structure of NNs is planned as a layered network and an interconnected network. The layered networks

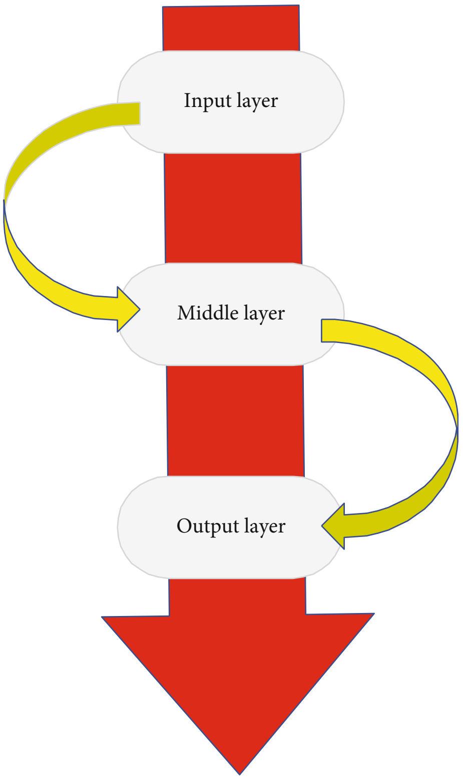


Figure 3: Schematic diagram of the layered network.

usually include input layer, middle layer, and output layer [22]. Figure 3 shows the schematic diagram of the layered network.

The layered networks can also be subdivided into: simple forward network, forward network with feedback signals, and forward network connected between layers. BP is one of the most widely used ANN models at present. It usually has three or more layers of multilayer NNs, each of which has many neurons. BPNN is a network of multilevel feedforward, which is trained according to the supervised learning method and error backpropagation algorithm [23]. Figure 4 shows the structure diagram of the BPNN model with only one middle layer.

The middle layer is the characteristic space, and the number of nodes is the dimension of the characteristic space. In BPNN, neurons receive learning mode. Any neurons on the left are linked to any neurons on the right. The activation value of neurons is transmitted to the output layer through the middle layer. The output feedback of neurons in the output layer obeys the basic principle of reducing the difference between the expected output value and the actual output value. It feeds back to each connection element through the hidden layer and the output layer, so it is also known as the "error backpropagation algorithm." With the continuous adjustment of connection weight, the error rate of the input mode response is reduced [24].

The number of nodes is n in the input layer of the program, the number of nodes in the middle layer is l, and the number of nodes in the input layer is m. In addition, the weight needs to be set. The weight from the input layer to the middle layer is W_{ij}, the weight from the middle layer to the output layer is W_{jk}, the bias value from the input layer to the middle layer is a_j, and the bias value from the middle layer to the output layer is b_k, and the learning rate is η. The incentive function is set $g(x)$ as a logarithmic S-type activation function, as shown in

$$g(x) = \frac{1}{1+e^{-x}}. \tag{28}$$

The output of middle layer is H_j:

$$H_j = g\left(\sum_{i=1}^{n} W_{ij}\chi_i + a_j\right). \tag{29}$$

The output of output layer is O_k:

$$O_k = \sum_{i=1}^{l} H_j w_{jk} + b_k. \tag{30}$$

$Y_k - O_k = e_k$, when Y_k is the expected output, the error calculation is shown in

$$E = \frac{1}{2}\sum_{k=1}^{m} (Y_k - O_k)^2. \tag{31}$$

If the error is minimal and the minimum value is min E, the weights are updated from the middle layer to the output layer and from the input layer to the middle layer by the method of gradient descent. The principle of error adjustment is to reduce the error value, which means that the weight correction of each layer should change in positive proportion to the negative gradient formed by the difference. The weight update value from the middle layer to the output layer is shown in

$$\frac{\partial E}{\partial w_{jk}} = \sum_{k=1}^{m} (Y_k - O_k)\left(\frac{-\partial Ok}{\partial w_{jk}}\right) = -e_k H_j. \tag{32}$$

The weight is $w_{jk} + \eta e_k H_j$ from the middle layer to the output layer.

The updated weight from the input layer to the middle layer is shown in

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial H_j} \cdot \frac{\partial H_j}{\partial w_{ij}} = -\sum_{k=1}^{m} w_{jk} e_k * H_j\left(1 - H_j\right) x_i. \tag{33}$$

The updated weight from the input layer to the middle layer is shown in

$$w_{ij} + -\eta \sum_{k=1}^{m} w_{jk} e_k * H_j\left(1 - H_j\right) x_i. \tag{34}$$

According to the above methods, the updated bias value

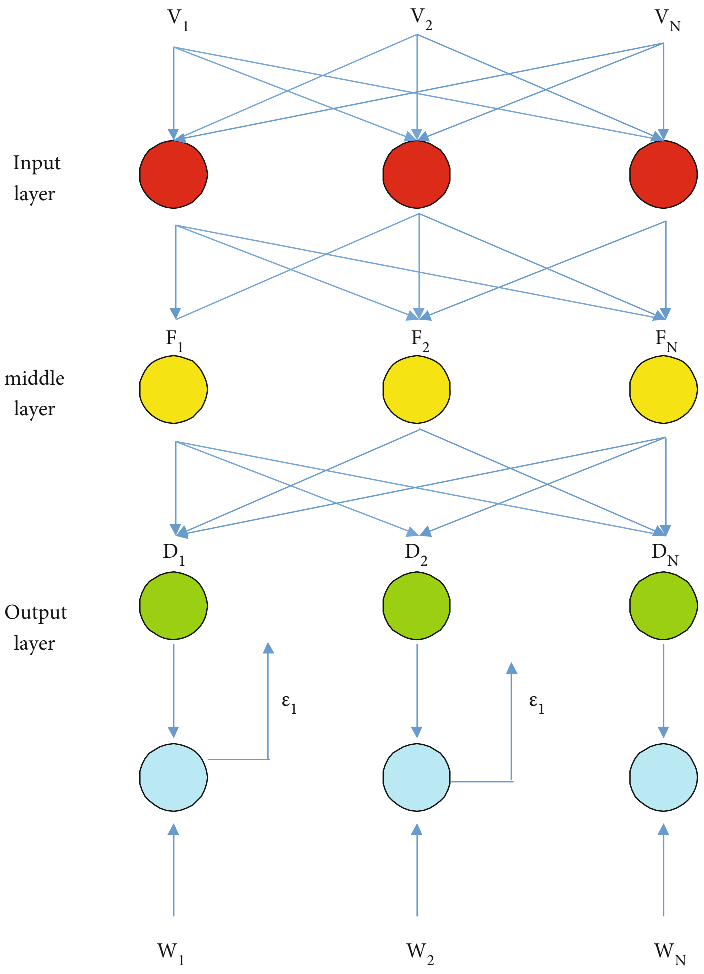


Figure 4: Structure diagram of the BPNN model with only one middle layer.

b_k from the input layer to the middle layer is shown in

$$b_k = b_k + \eta e_k. \tag{35}$$

The updated bias value a_j from the output layer to the middle layer is shown in

$$a_k = a_k + \eta H_j\left(1 - H_j\right) \sum_{k=1}^{m} w_{jk} e_k. \tag{36}$$

The input layer propagates backward to obtain the actual output. Compared with the expected output value, iteration stops if it reaches the accuracy of meeting requirements of the error function; if it is not achieved, it continues to update the weights of each layer until the accuracy required by the error function is reached.

The iterative algorithm must converge. The sequence of X_k converges to a certain minimum point $X_{\min}$. The equation is shown in

$$\lim_{k \longrightarrow \infty} \left| X_k - X_{\min} \right| = 0. \tag{37}$$

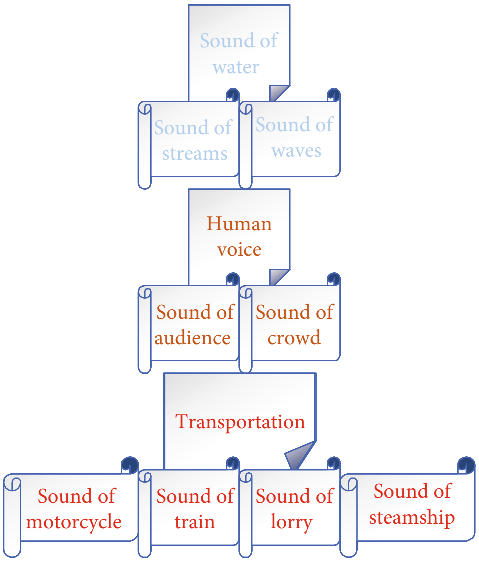


Figure 5: Schematic diagram of easily confused sound effects.

If the iterative sequence can converge to $X_{\min}$ by its starting point being close to the minimum point, it is called a local convergence algorithm, which is constrained by the minimum point. If any starting point produces an iterative

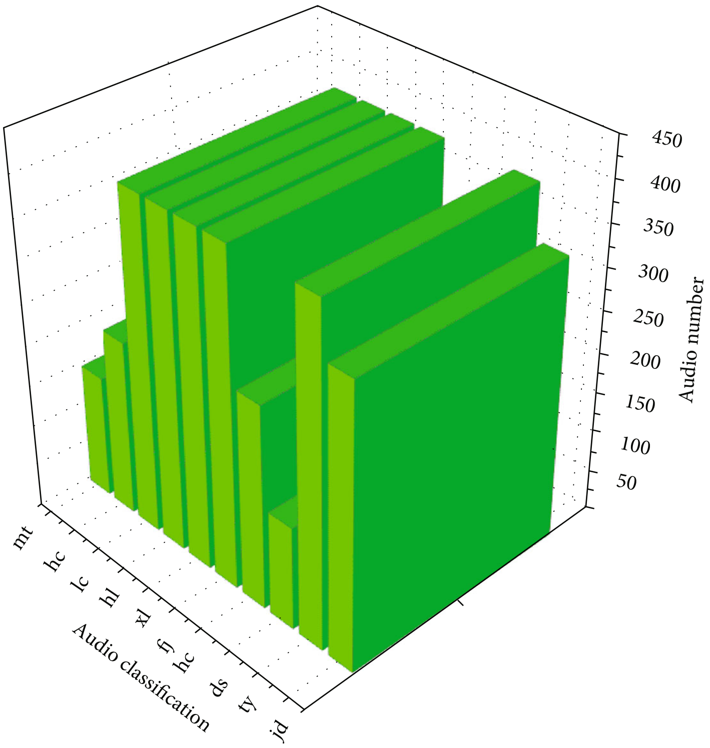


Figure 6: The number of HHM training samples.

sequence that can converge to $X_{\min}$, it is called a global convergence algorithm.

Using the iteration optimization algorithm, only the calculated iteration point X_k is understood, and the optimal solution $X_{\min}$ is not known. Therefore, it is necessary to judge when the iteration should end based on the information provided by the known iteration point. The termination condition is usually shown in equation (38):

In equation (39), determine whether the error is less than a predefined value.

$$|x_{k+1} - x_k| \le \varepsilon_1, \tag{38}$$

$$|f(x_{k+1}) - f(x_k)| \le \varepsilon_2. \tag{39}$$

In equations, there is the absolute error of two iterations. In some cases, the minimum relative error is required to judge the termination.

$$\frac{|x_{k+1} - x_k|}{|x_k|} \le \varepsilon_1, \tag{40}$$

$$\frac{|f(x_{k+1}) - f(x_k)|}{f(x_k)} \le \varepsilon_2. \tag{41}$$

There will also be cases of calculating the gradient mode. When the specified value range is reached, the iteration will be terminated, as shown in equation (41):

$$\Delta|f(x_{k+1})| \le \varepsilon \tag{42}$$

The quality of a NN design depends on the accuracy and the training time of the network. The construction of the BPNN determines the structure of the BPNN according to the characteristics of the input and output data of the system. The characteristic parameters have a total of 26 dimensions. There are 15 types of sound to be classified. Some problems can be solved with a single-layer network with a nonlinear activation function. Considering that the adaptive linear network can also be solved, and the correct rate of solving the problem with only a single-layer nonlinear function will not be too high, the number of layers must be increased to achieve better training accuracy. To improve the accuracy of network training, increasing the number of layers can further improve the accuracy rate and reduce the error. The experiment considers appropriately increasing the number of network layers without increasing the complexity of the network. After repeated simulation and training, it is finally confirmed that the selection of the BPNN training consists of 26-13-15. The input layer consists of 26 neurons, the middle layer consists of 13 neurons, and the output layer consists of 15 neurons. The 15 sound effects are reclassified to verify the recognition rate of the NN training model for confusing sound effects. The schematic diagram of easily confused sound effects is shown in Figure 5.

3. Analysis of the Simulation Results of Sound Feature of the Model

3.1. Simulation Results of Sound Feature Parameters Based on Hidden Markov. 10 sound effects are selected for hidden Markov modeling, namely, street crowd sound jd, stadium crowd sound ty, TV program sound ds, train sound hc1,

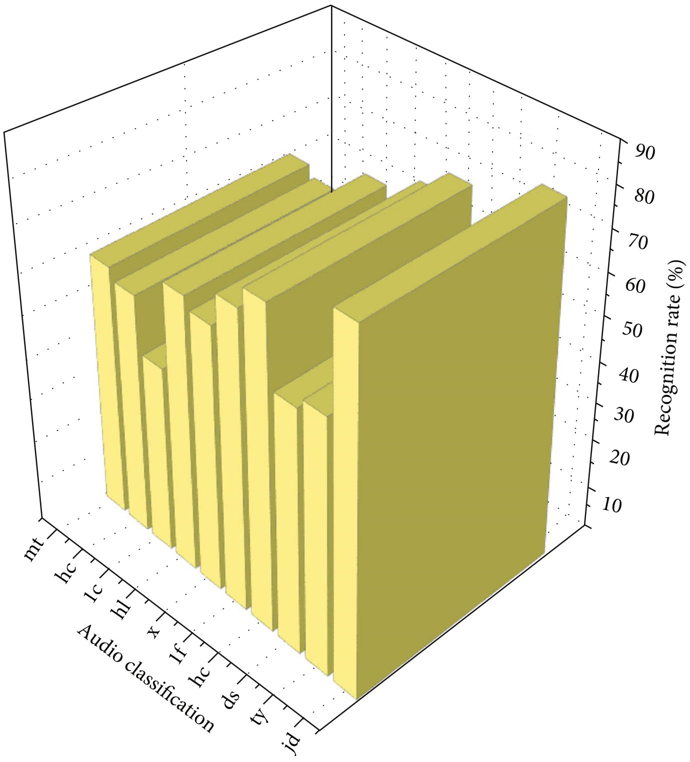


Figure 7: Recognition rate of each audio category.

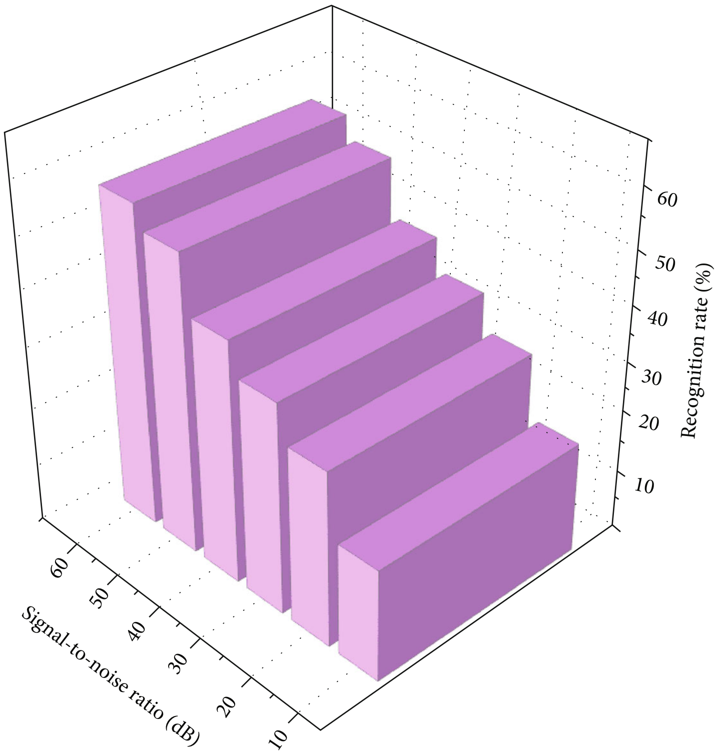


Figure 8: Test recognition rate under each SNR.

aircraft sound fj, stream sound xl, wave sound hl, ship sound lc, truck sound hc2, and motorcycle sound mt. The number of HHM training samples is shown in Figure 6.

Each frame parameter of audio file data, the first-order difference and the second-order difference of the 8th order, and short-term energy, and the short-term average

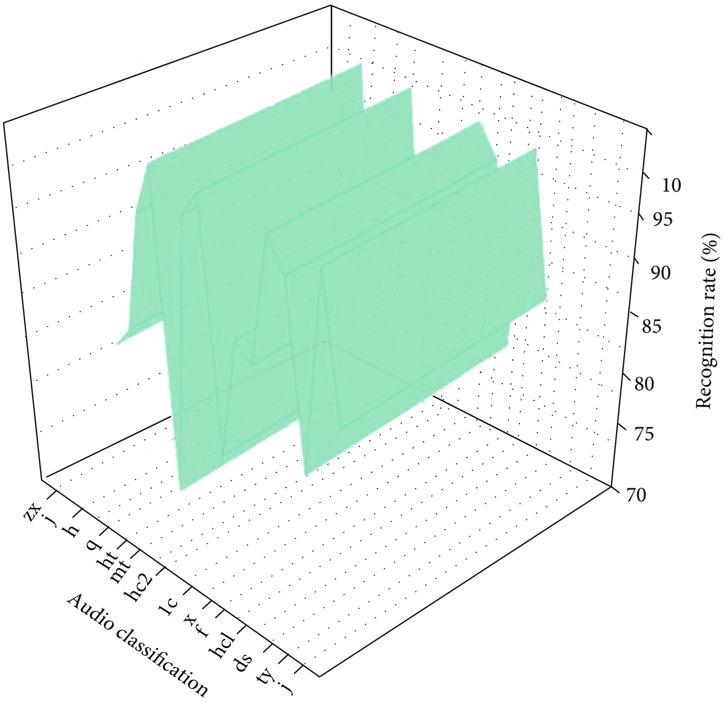


Figure 9: Accuracy of recognition of each category.

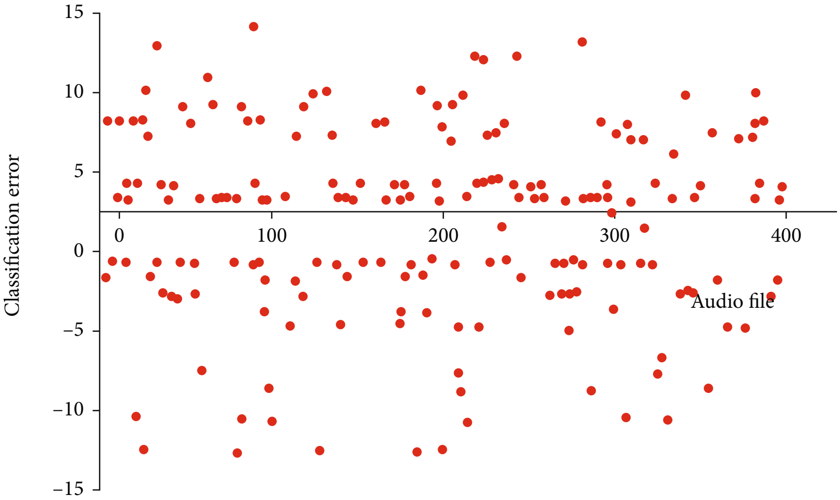
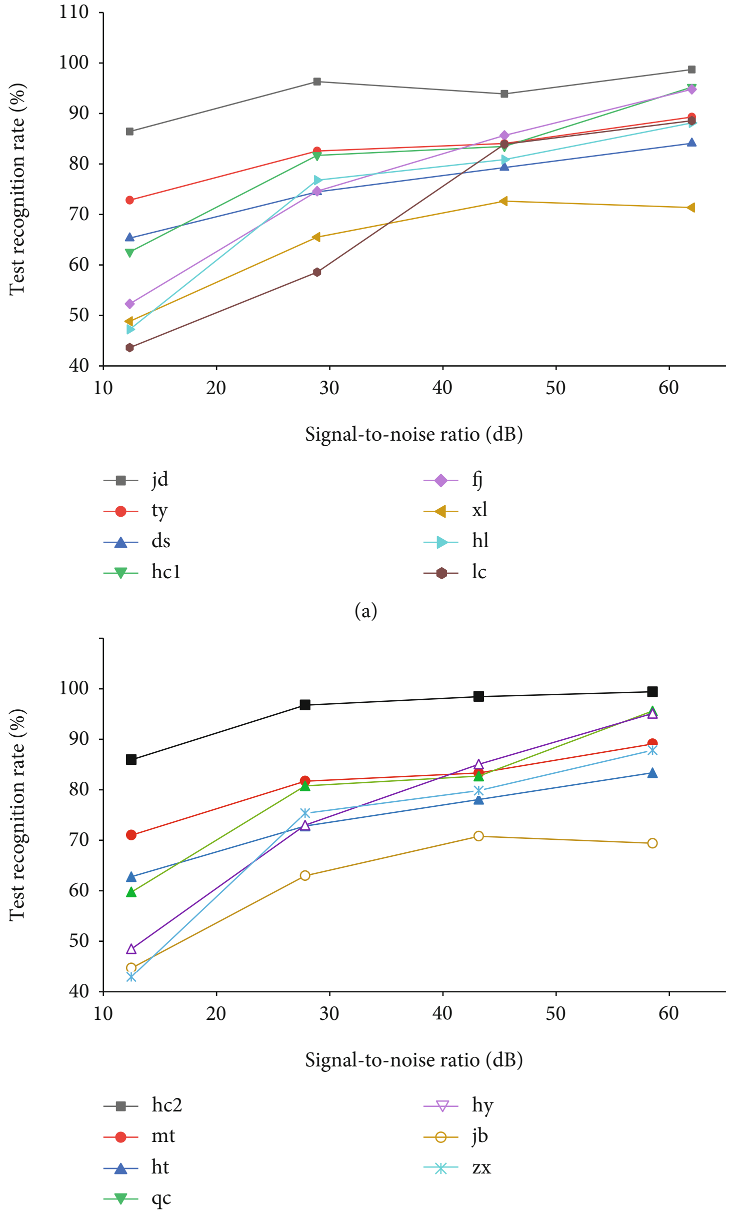


Figure 10: Scattered distribution of the actual classification and predicted classification errors of 100 to 500 audio clips randomly selected.

zero-over rate are extracted. A total of characteristic parameters are used as observation symbols in this experiment. In this experiment, the HHM is established for each classification. The recognition rate of each category is shown in Figure 7.

Among them, the recognition effect of stadium crowd voice and train sound is relatively ideal, and the rest of the results are unsatisfactory. There may be more confusing elements in the sound of vehicles in various categories, and the recognition effect is not good. The recognition rate of ships sound is less than 50%, which may be due to the sound of ocean waves has multiple characteristics, resulting in a low recognition rate. The experiment added Gaussian white noise that simulates the real environment to the sample to verify the anti-interference ability of the HHM. The signal-to-noise ratio (SNR) of the original tested sound materials was higher than 75 dB. After adding Gaussian white noise, the SNR was 10, 20, 30, 40, 50, and 60 dB, respectively. The comparison of test recognition rate is shown in Figure 8.

In Figure 8, as the SNR decreases, the recognition rate of the HHM has a significant downward trend, indicating that the anti-interference ability of this model is insufficient. To sum up, the Hidden Markov training model is not very suitable for describing sound materials containing more complex content, nor can it meet the needs of applying onomatopoeia materials.



(a)

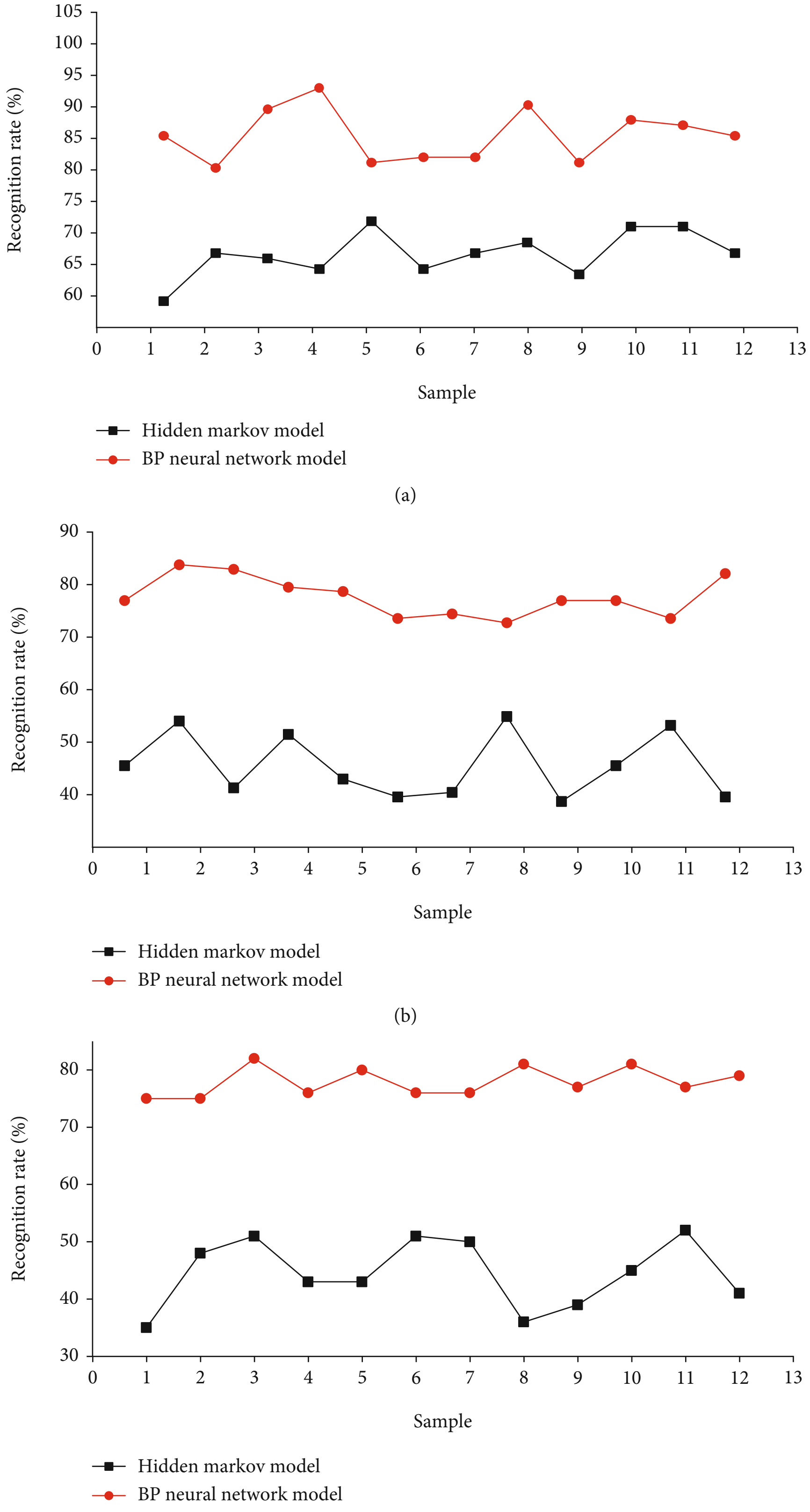
(b)

Figure 11: Test recognition rate of BPNN models in different SNR environments. (a) Test recognition rates of 8 sound categories in different SNR environments. (b) Test recognition rates of 7 sound categories in different SNR environments.

3.2. Simulation Results of Sound Feature Based on BPNN. The selection of the BPNN training consists of 26-13-15. The input layer consists of 26 neurons, the middle layer consists of 13 neurons, and the output layer consists of 15 neurons. A total of 15 typical sound materials are set, and five sounds are added to the above set: sound of children ht, sound of meeting hy, sound of footstep jb, sound of bicycle zx, and sound of car qc. The sound type can be divided into vehicle sound, human voice, and water sound. The recognition accuracy of each category is shown in Figure 9.

As shown in Figure 9, except for the low probability of children's voice being recognized, the recognition rate of all categories is more than 80%, of which human voice is more complex and should be further subdivided. The voice recognition rate of confusing stadiums has reached 100%, and the voice recognition rate of street can reach 90%. The recognition rate of conference sound is better than that of TV human voice, reaching 95.56%. The recognition rate of footsteps is better, and the rate of vehicles is relatively high. The recognition rate of cars, trains, and planes has reached 100%, but the recognition rate of ship sounds is not ideal, only 80.36%, which may be caused by the coincidence of the waves with certain characteristics. The recognition rate of sound of water is about 90%, which is ideal. To sum up, the recognition rate of confusing sound effects is relatively high, which is higher than 90% on average.

In real, sound may be damaged for various reasons, making it impossible to use later. Replaceable sound



(a)

(b)

(c)

Figure 12: Specific performance of the model in each scenario. (a) Specific performance of two algorithms under a single sound source. (b) Specific performance of two algorithms under a combined sound source. (c) Specific performance of two algorithms under a complex sound source.

resources can be found to compensate. Figure 10 shows the scattered distribution of the actual classification and predicted classification errors of 100 to 500 audio clips randomly selected.

The experiment is not very practical, and the characteristics of the sound material are not the same in the same classification. For example, the sound of the car engine, the sound of the door, the sound of airplanes taking off and landing, the sound of bicycle chain and bell, and other sounds are all distributed in the sound of the car. In the same scene, there are also new sounds replacing the original sound. Therefore, only use onomatopoeia to replace damaged materials is considered. Figure 11 shows the test recognition rate of BPNN models in different SNR environments.

Figure 11 indicates that the introduction of new sound materials by BPNNs does not have much impact on the original sound effects. Although it has a certain impact on the recognition rate, it is not easy to accurately recognize whether the new sound resources are complex or not. The BPNN has a good anti-interference performance when the SNR of the sound material is greater than 30 dB.

3.3. Comparison and Analysis of HMM and BPNN. From the previous section, it shows that the recognition result of the HMM is only about 60%, the accuracy of the BPNN algorithm is higher, and the average recognition rate reaches 91%. Because the previous selection of sound materials was randomly selected, the recognition effect was observed by selecting 3 sets of exactly the same sound materials to simulate in the two algorithm models. Figure 12 shows the specific performance of two algorithms under a single sound source, combined sound source, and complex sound source.

As Figure 12 indicates, the accuracy of HMM recognition is not as efficient as that of BPNN model. The reason may be that the HMM needs to cooperate with supervise learning algorithms to play a better role, but the BPNN model does not. Moreover, the HMM is relatively stable in the single sound source scenario, but the stability is greatly reduced when the sound source situation is slightly complicated.

In the BPNN model, the recognition effect of the single sound source, combined sound source, and compound sound source is better than that of the HMM, and the stability is relatively better. To sum up, the BPNN model has more advantages than the HMM for the recognition of sound effect materials. The BPNN model can achieve a higher recognition rate in a shorter training time and has a better generalization and reasoning ability and good performance of tamper resistance.

4. Conclusion

At present, the workload of sound effect classification by manual listening is large and cumbersome. Therefore, it is urgent to study the automatic classification of sound effect materials and improve work efficiency. To improve the accuracy of sound recognition, two algorithm models are established to automatically identify and compare sound materials, which are the HMM and BPNN models. First, the HMM is established, and the sound material is randomly selected as the test sample. The comparison between the expected classification and the actual classification is tested, and the recognition rate of each classification is obtained. The final average recognition rate is 61%. The anti-interference characteristics of the hidden Markov training model are tested under 6 types of SNR environment, and the recognition rate of the training model has a significant downward trend with the decrease of the SNR. Additionally, the BPNN model is established, and 200 training experiments of BPNN are carried out. The training model with the highest average recognition rate is selected as the final model in the experimental training. The average recognition rate of the final model is higher than 90%. It stimulates the expression ability and stability of the trained model after introducing new samples. And the tamper-interference performance of the model has been tested in different SNR environments. The performance test results are good, and only the recognition rate of complex sound types of individual sound sources has decreased. Finally, the accuracy of the HMM established in the experiment is not as high as that obtained by BPNN. Therefore, the BPNN training method has more advantages, and the automatic classification of sound effects can better meet the needs of practical applications, facilitate the work of the majority of audio workers, and provide a good theoretical basis for the automatic identification and classification of audio materials in the future. Due to some limitations, it needs to be further developed and improved in combination with practical applications so that it can be used better.

Data Availability

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest

The author declares that he/she has no conflicts of interest.

References

[1] M. Bensimon, S. Greenberg, and M. Haiut, “Using a low-power spiking continuous time neuron (SCTN) for sound signal processing,” *Sensors*, vol. 21, no. 4, p. 1065, 2021.

[2] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, “PANNs: large-scale pretrained audio neural networks for audio pattern recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, no. 12, pp. 2880–2894, 2020.

[3] S. Kwon, “A CNN-assisted enhanced audio signal processing for speech emotion recognition,” *Sensors*, vol. 20, no. 1, p. 183, 2020.

[4] H. Kwon, H. Yoon, and K. W. Park, “Acoustic-decoy: detection of adversarial examples through audio modification on speech recognition system,” *Neurocomputing*, vol. 417, no. 12, pp. 357–370, 2020.

[5] R. Haeb-Umbach, J. Heymann, L. Drude, S. Watanabe, M. Delcroix, and T. Nakatani, “Far-field automatic speech

recognition," *Proceedings of the IEEE*, vol. 109, no. 2, pp. 124–148, 2021.

[6] Q. Yu, Y. Yao, L. Wang, H. Tang, J. Dang, and K. C. Tan, "Robust environmental sound recognition with sparse key-point encoding and efficient multispike learning," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 2, pp. 625–638, 2021.

[7] Y. Jin, B. Wen, Z. Gu et al., "Deep-learning-enabled MXene-based artificial throat: toward sound detection and speech recognition," *Advanced Materials Technologies*, vol. 5, no. 9, p. 2000262, 2020.

[8] G. Li, Y. Xiong, Q. Du, Z. Shi, and R. S. Gates, "Classifying ingestive behavior of dairy cows via automatic sound recognition," *Sensors*, vol. 21, no. 15, p. 5231, 2021.

[9] L. Lhoest, M. Lamrini, J. Vandendriessche et al., "MosAIc: a classical machine learning multi-classifier based approach against deep learning classifiers for embedded sound classification," *Applied Sciences*, vol. 11, no. 18, p. 8394, 2021.

[10] F. Demir, D. A. Abdullah, and A. Sengur, "A new deep CNN model for environmental sound classification," *IEEE Access*, vol. 8, no. 4, pp. 66529–66537, 2020.

[11] Y. Zhang, J. Zeng, Y. Li, and D. Chen, "Convolutional neural network-gated recurrent unit neural network with feature fusion for environmental sound classification," *Automatic Control and Computer Sciences*, vol. 55, no. 4, pp. 311–318, 2021.

[12] R. Catanghal Jr., "Sound detection for study room monitoring and evaluation," *International Journal of Applied Science and Engineering*, vol. 18, no. 4, pp. 1–4, 2021.

[13] G. Gosztolya, "Using the fisher vector representation for audio-based emotion recognition," *Acta Polytechnica Hungarica*, vol. 17, no. 6, pp. 7–23, 2020.

[14] S. Cunningham, H. Ridley, J. Weinel, and R. Picking, "Supervised machine learning for audio emotion recognition," *Personal and Ubiquitous Computing*, vol. 25, no. 4, pp. 637–650, 2021.

[15] S. Zhao, Y. Ma, Y. Gu et al., "An end-to-end visual-audio attention network for emotion recognition in user-generated videos," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34no. 1, pp. 303–311, New York, New York, USA, 2020.

[16] D. Jones and L. Sun, "Search for continuous gravitational waves from Fomalhaut b in the second advanced LIGO observing run with a hidden Markov model," *Physical Review D*, vol. 103, no. 2, article 023020, 2021.

[17] T. Xia and X. Chen, "A discrete hidden Markov model for SMS spam detection," *Applied Sciences*, vol. 10, no. 14, p. 5011, 2020.

[18] A. Melatos, L. M. Dunn, S. Suvorova, W. Moran, and R. J. Evans, "Pulsar glitch detection with a hidden Markov model," *The Astrophysical Journal*, vol. 896, no. 1, p. 78, 2020.

[19] C. Guo, M. Zhang, and H. Chen, "Suitability of low-field nuclear magnetic resonance (LF-NMR) combining with back propagation artificial neural network (BP-ANN) to predict printability of polysaccharide hydrogels 3D printing," *International Journal of Food Science & Technology*, vol. 56, no. 5, pp. 2264–2272, 2021.

[20] D. Zhang and S. Lou, "The application research of neural network and BP algorithm in stock price pattern classification and prediction," *Future Generation Computer Systems*, vol. 115, no. 2, pp. 872–879, 2021.

[21] X. Wang, J. Yuan, and B. Wang, "Prediction and analysis of PM2. 5 in Fuling District of Chongqing by artificial neural network," *Neural Computing and Applications*, vol. 33, no. 2, pp. 517–524, 2021.

[22] Y. Zhang, J. Tang, R. Liao et al., "Application of an enhanced BP neural network model with water cycle algorithm on landslide prediction," *Stochastic Environmental Research and Risk Assessment*, vol. 35, no. 6, pp. 1273–1291, 2021.

[23] R. K. Pandey, T. Y. Lin, and P. C. P. Chao, "Design and implementation of a photoplethysmography acquisition system with an optimized artificial neural network for accurate blood pressure measurement," *Microsystem Technologies*, vol. 27, no. 6, pp. 2345–2367, 2021.

[24] F. Yang, J. Mou, Y. Cao, and R. Chu, "An image encryption algorithm based on BP neural network and hyperchaotic system," *China Communications*, vol. 17, no. 5, pp. 21–28, 2020.