Research on Music Classification Based on MFCC and BP Neural Network

LiuYongchun and Song Hong
School of Automation & electronic information
Sichuan University of Science & Engineering
Zigong, Sichuan Province, P.R.China
028lyc@163.com

Yang Jing
School of Foreign Language
Sichuan University of Science & Engineering
Zigong, Sichuan Province, P.R.China
hnetyj@qq.com

Abstract—**Because of the diversity and uncertainty of music, the classification rate and accuracy are both lower for the traditional classification methods in the large-scale music classification application. A based on BP neural network (BPNN) music classification method proposed in this paper can improve this performance, which extracts the feature parameters of music through mel frequency cepstrum coefficient(MFCC) firstly, and then the BPNN is used to train feature signals and establish the optimal classifier model, finally classifies the test music dataset. The average classification accuracy rate is up to 90.2%, and higher 7% than the HMM classification method by simulation experiments for the folk, classical, rock and pop different types of music, therefore, the results show that the BPNN is a quite effective music type classification method.**

Keywords-BP neural network; MFCC feature extraction; music classification; hidden Markov model

I. INTRODUCTION

Due to the complexity of music itself, the ambiguity of music type definition and the limited understanding for human auditory perception properties, people find it is very difficult to query their demand target music when faced to the massive digital music resources growing rapidly on the Internet and in the Digital Library. In order to make the query be convenient, these resources often need to be organized and classified according to some rules. Moreover, music automatic classification is the key technology to realize the fast, efficient retrieval, in the same time, is an important part of multimedia information analysis and retrieval technology based on content, which must promote the researches on music similarity analysis, music recommendation system, automatic music transcription system and other related fields [1].

Music classification is essentially a pattern recognition process, mainly including the following several function models, the speech signal preprocessing, feature extraction, classifier training and classification test. A lot of research results have been achieved in this area, such as the audio classification based on rules, pattern matching, hidden Markov models (HMM) and so on, but these methods have their own shortcomings. For the first method, its operation is simple, but it is only applicable to recognize these music types with simple features, such as mute, which is difficult to satisfy these classification applications with complex and multi features. The second method, it first needs to build a standard pattern for each audio type, and then compares the input pattern with the standard pattern, which has large amount of calculation, low classification accuracy. The HMM has the strong modeling ability for the dynamic time series, small amount of calculation, but classification decision ability is poor, needs the prior knowledge of statistics and other defects. Because of the diversity, uncertainty and huge amount of music, we have to select the classifier according to the particularity of music classification, BPNN can simulate human neuronal activity principle, have the self-organization, adaptability, and continuous learning ability, and can approximate any nonlinear functions, parallel process information, have strong fault tolerance ability and many other advantages, which is especially suitable for these applications like music classification that is difficult to describe using algorithm and has a large number of samples for the algorithm study. In this paper, first adopting the MFCC method extract music features, and then using BPNN model as a classifier, extracting several samples from the same audio, and finally using the voting method to determine the music type for these samples recognition results [2].

II. MFCC FEATURES EXTRACTION

The music features are divided into time domain and transform domain features, as the human ear different feelings to the different speech signal mainly depend on their short-time amplitude spectrum, especially their resonance peak positions and their width, in addition, the time domain features calculation is simple, but not suitable for describing the amplitude spectrum features. MFCC reflects a kind of subjective feeling of people to the level of sound, the computation amount is small extracting the cepstral features parameters from the Mel scaling frequency domain, which has been widely used in the field of speech signal processing. The MFCC coefficients used as audio classification features, can improve the classification accuracy. The MFCC feature extracting processes of music signals are as the following:

- The input speech signals are divided into frames, window, and then transformed. The incoming speech signal is divided into frames, multiplied with window function, and then calculated the Fourier transform, so as to obtain frequency spectrum distribution information. DFT of the speech signal is given by:

$$X(k)=\sum_{n=0}^{N-1}x(n)e^{-j\frac{2\pi nk}{N}}, \quad 0\le k\le N-1 \quad .!$$

Where x(n) is input signal, N expresses the points number of DFT.

- Computing the spectrum amplitude square yields the energy spectrum.
- Make the energy spectrum pass through a triangle filter bank with Mel scale.

Define a filter bank with M (M=100) filters, and adopt the triangle filters, which center frequencies are respectively given by: f(m),m=1,2,3,…M.

- Calculate the logarithm of energy for each filter bank output, we obatain

$$S(m)=\ln(\sum_{k=1}^{N-1}\left|X(k)\right|^{2}H_m(k)), \quad 0\le m\le M-1\ .$$

Where $H_m(k)$ is the frequency response of the triangle filter.

- yield the MFCC coefficients by Discrete cosine transform.(DCT).

$$C(n)=\sum_{m=0}^{M-1}S(m)\cos(\pi n(m-0.5/m)), \quad 0\le n\le N-1 \ \cdot$$

Usually select the number of MFCC coefficient as 20-30, often without using the 0 order coefficient, because it reflects the spectrum energy, which is called energy coefficient in the general recognition system without be used as cepstrum coefficient. Therefore, we select 24 orders cepstrum coefficients in this paper [3].

III. MUSIC CLASSIFICATION BASED ON BPNN

A. BP Neural Network

BP neural network is a multilayer feedforward neural network, which main characteristic is that the signals always forward transfer, the errors backward transfer. In the forward transferring process, the input signals are processed layer by layer from the input layer and through the hidden layers and until the output layer, meanwhile, the state of the neuron of each layer only affects the neuron state in the next layer. If the outputs of the output layer are not the expected values, then backward transfer and adjust the network weights and the thresholds according to the predicted error, so that the predicted output values continuously approximate the expected output values. The topology of the BPNN structure is shown in Figure1. Where $X_1,X_2,\dots,X_n$ represent the input values of the BPNN, $Y_1,Y_2,\dots,Y_m$ represent the predicted values, and ω_{ij},ω_{jk} represent the network weights. The BPNN can be considered as a nonlinear function, the input values and the predicted values represent the independent variables and dependent variables respectively. When the number of the input nodes and the output is n and m respectively, and then the BPNN expresses a functional mapping relationship with n independent variables and m dependent variables [4].

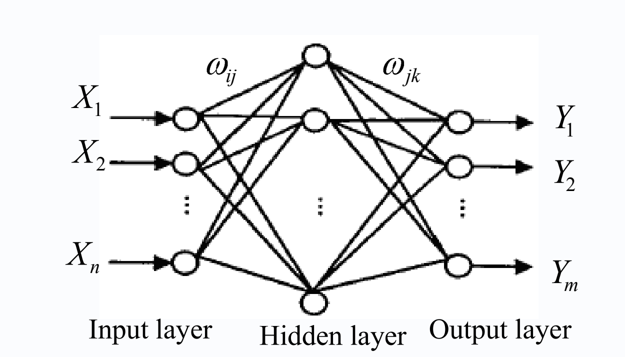

Figure 1. BP neural network structure.

BPNN classifier first needs to train, so that the network has associative memory and predicting abilities. The training steps are as the following:

- Initialize the network. According to the input and output sequences (X,Y) of the system, determine the number n of the input layer nodes, number l of the hidden layer nodes and number m of the output layer nodes, and then initialize the ω_{ij},ω_{jk} and the thresholds a of the hidden layer and b of the output layer, set the learning rate and the neuron excitation function.
- Calculate the output of the hidden layer. According the input vector X, the weight ω_{ij} between the input layer and the hidden layer, and the threshold a, yield the output H of the hidden layer.

$$H_j=f(\sum_{i=1}^{n}\omega_{ij}x_i-a_j) \quad j=1,2,\dots,l \quad . \ .$$

Where the excitation function is defined as

$$f(x)=\frac{1}{1+e^{-x}} \quad . \ \cdot$$

- Calculate the output of the output layer. According to the above H ,the weight ω_{jk} and threshold b, compute the output O of the BPNN.

$$O_k=\sum_{j=1}^{l}H_j\omega_{jk}-b_k \qquad k=1,2,\dots,m \quad . \ .$$

- Calculate the error. According to the predicted O and the expected output Y, compute the error e.
- Update the weights. According to the predicted error e, update the weights ω_{ij},ω_{jk}.

$$\omega_{ij} = \omega_{ij} + \eta H_j(1-H_j)x(i)\sum_{k=1}^{m}\omega_{jk}e_k \tag{7}$$

$i = 1,2,\ldots,n$; $j = 1,2,\ldots,l$

$$\omega_{jk} = \omega_{jk} + \eta H_j e_k \qquad j = 1,2,\ldots,l;\ k = 1,2,\ldots,m \tag{8}$$

Where η is the learning rate.

- Update the thresholds. According to the predicted error e ,update the thresholds a and b .

$$a_j = a_j + \eta H_j(1-H_j)\sum_{k=1}^{m}\omega_{jk}e_k \tag{9}$$

$j = 1,2,\ldots,l$

$$b_k = b_k + e_k \qquad k = 1,2,\ldots,m$$

- Determine whether the algorithm iteration is the end, if not, then return to step 2.

B. Music Classification Based on BPNN

We select four types of music, for example, folk, classical music, rock and pop, using the BPNN realize the classification for the music. First, using the MFCC extract 500 groups of 24 dimensional feature signals, as shown in Figure 2. The classification algorithm based on BPNN includes three functional modules, that is, constructing BPNN model, training network and classifying. When construct the BPNN model, according to the input and output data features, define the BPNN structure, because the music feature input signals has 24 dimensions, the music signals waiting to be classified has 4 classes, set the input nodes number equal 24, the hidden nodes number equal 25, and the output nodes number equal 4. There are 2000 groups of music signals, randomly select 1500 groups of data as training data to train the BPNN, and the other 500 groups of data will used as the testing set to test the classification ability, finally, using the trained BPNN, classify or recognize the data set [5].

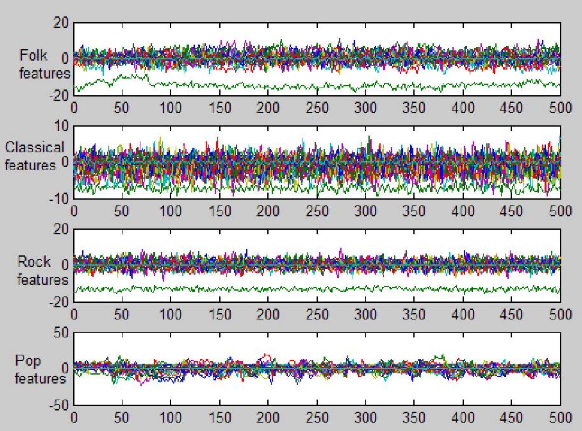

Figure 2. Music features

IV. EXPERIMENT RESULTS AND ANALYSIS

Adopt the MATLAB 7.13 development software as the experiment platform, program and realize music signals classification based on BPNN.

A. Data Source and Preprocessing

a) data source

First using the MFCC extract the four different class music signals features, and identify their classes respectively by 1,2,3,4. The extracted feature vector has 25 dimensions, in which the first dimension is the class identification, and the other 24 dimensions are the MFCC feature coefficients, at the same time, these vectors are respectively stored in data1.mat,data2.mat,data3.mat,data4.mat database files. And then put the four classes of music signals are combined into a group, according to the classes identification, set the expected output values for each group signals, for example, for identification 1, we can set the output vector equals [1 0 0 0].

b) Data Preprocessing

Before the neural network predict, data often need to be normalized, which purpose is to transfer all data into a number between [0,1], so as to eliminate the phenomenon that the predicting error is too big caused by the bigger difference of the order of magnitude in every dimension data. The common normalization method is the maximum minimum method, that is:

$$x_k = (x_k - x_{\min}) \Big/ (x_{\max} - x_{\min}) \tag{10}$$

Where $x_{\min}$ is the minmum value of the data sequence, $x_{\max}$ is the maximum. We can directly adopt the mapminmax function in MATLAB to realize normalization [6].

B. Experiment Results and Analysis

Using the trained BPNN classifies the four classes of music signals, and the experiment results and classification error are respectively as shown in Figure3 and Figure4. The results indicate that the music signals classifying algorithm based on BPNN has higher accuracy rate, the average correct classification rate is up to 90.74%, especially for the classical music, up to 100%. It shows adopting the classification method is quite reasonable and the performance is better. In addition, we also use the HMM algorithm to classify these music feature signals, which correct rate and the BPNN result are shown in TABEL 1. By comparing, we know the recognition ability of the BPNN is much better.

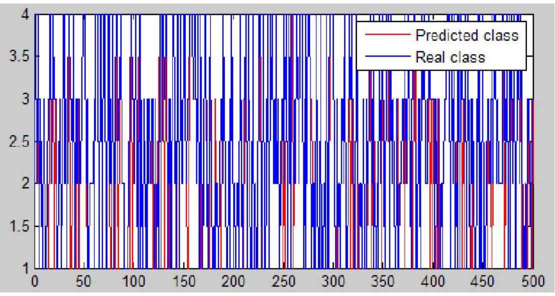

Figure 3. BPNN predicted result

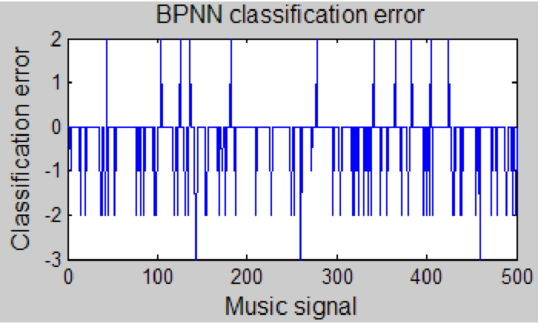

Figure 4. BPNN classification error

TABLE I. CLASSIFICATION CORRECT RATE (%)

Class of Music	**BPNN**	**HMM**
folk	84.5%	77.65%
classical	100%	95.38%
rock	85.2%	78.42%
pop	93.28%	88.31%

V. CONCLUSIONS

In this paper, first adopting the MFCC extracts the music features, and then using the BPNN algorithm recognizes the music classes, furthermore, simulation experiments indicate its performance is better. In the network training process, the BPNN uses the gradient correction method to update the weights and the thresholds from the negative gradient direction of the predicting error without considering the accumulation of the previous experience, which leads the learning process to converge slowly. In order to improve it, we can use the additional momentum method to solve the slow convergence speed problem. In addition, the BPNN learning rate η is the value in the [0,1], if η is much bigger, the weight modification is much greater, and then the network learning speed will more quickly. But the too big learning rate will cause oscillating in the weight learning processes, meanwhile, the too small rate will make the convergence speed slowly, so that the weight will be unstable. To improve this problem, we can use the variable learning rate, that is, set the rate to be appropriate big in the beginning learning process so that the network convergence speed is quickly, with the learning process continuing, the rate decreases, and then the network tends to be stable.

Acknowledgment

This research is supported by Artificial Intelligence Key Laboratory of Sichuan Province. (Grant No.2010RZ001)

References

[1] LI Hui-min, ZHANG Ren-jin. Research of vehicle license plate recognition based on BP neural network [J]. computer engineering and design, 2010,31 (3),619-621.

[2] G Tzanetakis , P Cook . Musical genre classification of audio signals [J] . IEEE Trans . on Speech and Audio Processing, 2002,10(5), 293 -302.

[3] ZhangYan, Tang Zhenmin, LiYanping, ZouYi. Research of music classification based on MFCC feature and HMM model[J]. Journal of NanJing normal university(Engineering and technology edition), Vol. 8, No. 4, Dec, 2008, 152-156.

[4] Tobias Herbig, Franz Gerl,Wolfgang Minker. Self-learning speaker identification for enhanced speech recognition[J]. Computer Speech & Language, Vol.26, No.3, Jun, 2012, 210–227.

[5] Toru Taniguchi, Mikio Tohyama, Katsuhiko Shirai.Detection of speech and music based on spectral tracking[J]. Speech communication, Vol.50, No.7, Jul,2008,547-563.

[6] Stefano Scanzio,Sandro Cumani,Roberto Gemello,Franco Mana,P. Laface. Parallel implementation of Artificial Neural Network training for speech recognition[J]. Pattern Recognition Letters, Vol.31, No.11, Aug, 2010, 1302-1309.