Research on timbre classification based on BP neural network and MFCC

Zhihang Meng¹
1 School of Modern Post(School of Automation), Beijing University of Posts and Telecommunications, Beijing ,100876, China
* e-mail: mengzhihang@bupt.edu.cn

Abstract. Music exists in all aspects of life. With the development of deep learning and neural network technology in recent years, music information retrieval has become an emerging field that has received widespread attention. In the past, musicians often relied on manual labeling to classify music. This method is not only time-consuming and labor-intensive, but also inaccurate. Although some researchers have tried to extract music features for automatic classification, it is difficult to extract appropriate audio features due to the interaction between the fundamental wave and harmonics of the music itself. In this paper, the Mel cepstrum coefficient is selected to extract the timbre characteristics, and the improved BP neural network is used to process the characteristics, which overcomes the previous problems and obtains a higher accuracy rate.

1. Introduction
Music can be found everywhere in life, and various music styles are formed by different timbre of disparate musical instruments, which has brought distinctive experiences and feelings to music fans. Therefore, a new research field has been established based on the classification and retrieval of music characteristics[1]. But since the style of a piece of music is often relevant to factors like rhythm, melody, timbre and others, and the understanding differences towards music of different artists have caused the instability of volume, tone and other features in the music. Thus, most of the present music classification and retrieval are achieved through text label information. In this way, the accuracy of the text label is highly demanded, but also the maintenance cost is high. The study suggests that the timbre differences of different musical instruments are not only more prominent, but also more stable in music. Therefore, it is feasible to adopt the method based on timbre retrieval. Literature [2] has classified different music genres by using Convolutional Neural Network (CNN), but hasn’t set different threshold values according to different music, which has contributed to differences in the accuracy of music classification among various genres. Non-negative matrix is used to decompose NMF and probabilistic latent components is utilized to analyze PLCA, which jointly contributes to estimate pitch in Literature [3-4], but it is difficult to identify sustained music. The method, short-term autocorrelation, is selected in Literature [5] to detect pitch, but the method is not universal since it is limited to an octave and the range is narrow. Literature [6] use the subharmonics to harmonics ratio (SHR) to estimate pitch, which effectively overcomes the problem of identifying sustained notes, but the frequency range can be recognized is narrow. Literature [7] combine BP neural network and Softmax regression model to identify it, but only single-tone audio can be identified in spite of the accurate recognition.

This paper has selected the Mel cepstrum coefficient (MFCC) to extract the timbre characteristics of notes, obtain the timbre features of different musical instruments, and then use the improved BP neural network to effectively identify the timbre features. It is by this means that the outcome can be relatively accurate and feasible in view of the shortcomings of the above model.

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd
2. Principles

2.1. Timbre extraction of different musical instruments
Timbre refers to distinctive characteristics of different sounds in expression of waveforms [8]. Different sounding bodies produce timbre differently on account of their various materials and structures.

It is the uniqueness of timbre that we distinguish people from things through sounds. Different vibrations can always be made up distinctive sounds, that is why timbre appears distinctively. Because of fundamental tone and overtone, musical instruments can produce diverse and wonderful sounds. Different timbre depends on various combinations of the two tones, although some musical instruments is similar to others in fundamental tone due to their own structure and materials, it is hard for non-professionals to differentiate these timbre differences. However, the characteristic differences are obvious in the spectrum. Therefore, the spectrum can be used to conduct effective and accurate analysis.

2.2. Conventional methods for extracting audio features
This paper adopts the transform domain features of audio signal for analysis since it is hard to effectively differentiate the timbre in time domain. The transform domain features of audio signals include frequency domain and cepstrum domain, and the former is comprised of Short-Time Fourier Transform (Short-Time Fourier Transform, STFT) [9], Constant Q Transform (CQT), etc. While cepstral domain is formed by Mel cepstrum coefficient (MFCC), etc. Considering the importance of the low frequency domain features of the music signal, the Short-Time Fourier Transform (STFT) is frequently used to found the music spectrum. This method can primely keep the low frequency band features. However, because the frequency of the Discrete Fourier Transform (DFT) is linearly separated, the frequency of the note doubles an octave (every 12 notes). This leads to some deficient identification of notes, also has a certain effect on the extraction of timbre. Like short-time Fourier transform, Constant Q transform (Constant Q transform) [10] is an important analysis tool for time-frequency, and its window length changes with frequency, which have effectively overcome the drawbacks of short-time Fourier transform. but too much attention to overtone is not ideal for extracting fundamental frequency with regard to timbre extraction. In line with the perceptual characteristics of human ears, the Mel cepstrum coefficient (MFCC) is extracted and it plays an important role in timbre recognition. The problems in the above methods can be effectively resolved by reasonably setting the filter. Therefore, this paper has selected the Mel cepstrum coefficient (MFCC) as the feature of timbre extraction.

3. Mel-Frequency Cepstral Coefficients
Mel-Frequency Cepstral Coefficients(MFCC) [11-12] is a feature extraction approach based on the nonlinear characteristics of human hearing. The resemblance between the extracted features and human hearing make it an ideal tool to represent the features of different instruments’ timbre. The study of auditory mechanism of human ears found that when two sounds with different loudness act on human ears, the frequency with higher loudness will affect the perception of frequency components with lower loudness, making them imperceptible. This phenomenon is called masking effect. The masking effect makes it easy for bass to mask treble, while the other way around would be extremely difficult. That is to say, for chords that emit sounds simultaneously, the highest tone on the top layer is easily recognized while the two sounds at a lower layer can hardly be identified. This would affect the accuracy of recognition. Therefore, the introduction of MFCC could help establish a set of band-pass filters based on density of the critical bandwidth in the frequency band from low frequency to high frequency to filter the input signal. After obtaining the logarithmic energy of the filter output, DCT is performed to obtain MFCC. Among them, converting the frequency of music to Mel frequency is the basis for extracting Mel cepstrum coefficients. The relationship between MFCC and frequency is given by Formula (1):

\[
\text{Mel}(f) = 1125 \times \left(1 + \frac{f}{700}\right)
\]  (1)
where $f$ represents the perceived frequency in the unit of $\text{Mel}$, also called the Mel-scale. Its relationship with linear frequency is shown in Fig 1:

![Fig 1. The relationship between mel cepstrum coefficient and frequency](image.png)

From the above figure, in the high frequency band, the Mel frequency gradually tends to be flat. If it is not processed, the effective information in the high frequency band will be lost. To extract MFCC, the signal needs to be segmented and pre-emphasized to highlight the high-frequency part of the signal. The formula is as follows:

$$s_i(n)i' = si(n) - \beta * s_i(n - 1) \quad (2)$$

Then, the energy spectrum is obtained by discrete Fourier transform using the following formula:

$$S_j(n) = \sum_{n=1}^{N} s_j(n)h(n)e^{-j2\pi k n/N}, 1 \leq k \leq K \quad (3)$$

where $K$ represents the length of discrete Fourier transform, $h(n)$ represents the analysis window with the sampling length of $N$. The hamming window was selected in this paper. Windowing would smooth the signal during the transformation and reduce the sidelobe size and spectrum leakage after Fourier transform. The hamming window function is as follows:

$$w(n,\alpha) = (1 - \alpha) - \alpha \cos(2\pi n/N), 0 \leq n \leq N - 1 \quad (4)$$

In the above formula, $\alpha$ is the coefficient (usually $\alpha = 0.46$), and the periodic energy spectrum of each frame is estimated as follows:

$$P_i(k) = \frac{1}{N} |S_i(k)|^2 \quad (5)$$

Then, determine the upper and lower limits of signal frequency using Formula (6):

$$M^{-1} = 700 \left( \exp\left( \frac{m}{1125} \right) - 1 \right) \quad (6)$$

The selected value expressed by Mel scale is converted to the frequency domain, thus the corresponding position of the filter is determined. Then take the logarithm of the vector after filtering, and perform discrete cosine transform (DCT):

$$C(n) = \sum_{m=0}^{N-1} s(m) \cos \left( \frac{\pi n(m - 0.5)}{M} \right), n = 1, 2, ..., L \quad (7)$$

From which the MFCC is obtained.

Since this feature does not depend on the nature of the signal, it doesn’t make any assumptions or restrictions on the input signal, and it uses the human ear’s hearing model as well. Therefore, this parameter has better robustness than the linear prediction cepstral coefficient (LPCC) based on the vocal tract model and more in line with the auditory characteristics of the human ear, which bring better timbre recognition performance.

4. **BP neural network**

4.1. **The principle of BP neural network**
BP (Back Propagation) network, short for back propagation neural networks, is a multilayer neural network of one-way transmission\cite{13-15}. As the basic concept, method of least squares utilizes the grad search technique, which has minimized error mean square of the actual output value and the expected output value of the network. And its learning process contains forward propagation and back propagation. In the course of forward propagation, the input information is processed by the hidden layer from the output layer and transmitted to the output layer. The state of the neurons in each layer will only affect the next layer. It has partly overcome the problem of local convergence produced by excessive learning rate. If the ideal output hasn’t been obtained in the output layer, the network will enter the state of back propagation, return the error signal along the original connecting path, and modify the weights of the neurons in each layer at once so as to minimize the error. The schematic of BP neural network is shown in fig 2.

\[ \text{The output layer is the part of the output result of the BP network, and the formula is:} \]
\[ o_k = f(\text{net}_k) \quad k = 1,2,...,l \quad (8) \]
\[ \text{net}_k = \sum_{j=0}^{m} w_{jk} y_j \quad k = 1,2,...,l \quad (9) \]

Since \( f(x) \) is the transfer function, in view of nonlinear feature of the extracted timbre, bipolar Sigmoid function is selected as the transfer function, so as to optimize the weights of each layer in the network and speed up the training pace, and the formula is as follows:

\[ f = \frac{1-e^{-x}}{1+e^{-x}} \quad (10) \]

The BP network employ the error \( E \) of the output result conducting the back propagation, then updates the weight value, and finally reaches optimal solution. The definition of output error \( E \) is as follows:

\[ E = \frac{1}{2} \sum_{k=1}^{l} (d_k - o_k)^2 \quad (11) \]

Unfolding the definition of the output error \( E \) to the hidden layer, the formula of the error \( E \) in the layer can be got as follows:

\[ E = \frac{1}{2} \sum_{k=1}^{l} \left[ d_k - f \left( \sum_{j=0}^{m} w_{jk} y_j \right) \right]^2 \quad (12) \]
Further upholding the above formula to the input layer, the formula of error $E$ in the layer is as follows:

$$E = \frac{1}{2} \sum_{k=1}^{l} \left( d_k - f \left( \sum_{j=0}^{m} w_{jk} f \left( \sum_{i=0}^{n} v_{ij} x_i \right) \right) \right)^2$$ \hspace{1cm} (13)

By constantly calculating errors and updating weights, the accuracy of the BP network will become higher and higher, and quickly reach a stable output value that is of good stability, accuracy and rapidity. The processing procedure is shown in Fig 3.

4.2. The improved method at timbre recognition

4.2.1. Space of programming sample

The audio data sets of various musical instruments that have been publicly available are either incomplete or too small. Therefore, it seems particularly significant to design a reasonable screening rule to construct the initial sample space of the network \( [16] \). The following illustrates it through a case study of constructing a piano timbre sample.

In the first place, the main timbre of the constructed sample should be confirmed. Many audios employ the way of duets in the currently available data sets, also transient harmonies of other instruments will arise in the course of playing. In this paper, the timbre has been classified into primary timbre, secondary timbre and negligible timbre, by detecting the duration of different timbre in music in line with the time scale. A threshold has been set in the net for negligible timbre, and the part that haven’t exceeded the threshold can be ignored directly, which has improved the training speed and accuracy of the network.

In the next place, the initial input vector space and its corresponding output value vector will be established after the main timbre being selected, which are based on the existing single timbre music period. Thereinto, the input vector contains $x$ and $y$, which is the weight of the primary timbre and the secondary timbre respectively. In this way, a new sample is established for training, which has expanded the training data set and improved the accuracy of the network.

![Fig. 3 Processing procedure of BP neural network](image-url)
4.2.2. Simulated Annealing Algorithm

In the initialization process of BP neural network, problems are easy come into being like slow learning pace and local optimal solution caused by unreasonable selection of initial weights. Therefore, Simulated Annealing Algorithm\textsuperscript{[17-18]} is used to optimize the initial weights so as to avoid the emergence of local optimal solution in selecting the initial weights. Firstly, the productive function is used to generate an initial value in the space of solution, and it is usually the method that can generate a new solution by simply transforming the current new solution that would be chosen, such as replacing and exchanging all or part of the elements that form the new solution. Secondly, the difference of objective function between it and its corresponding initial value is calculated by the method of incremental computation. Afterwards, whether the initial value is accepted or not is judged in accordance with the commonly selected Metropolis formula: if \( \Delta T < 0 \), \( S' \) will be the new current solution \( S \), otherwise the probability \( \exp(-\Delta T / T) \) accepts \( S' \) as the new current solution \( S \).

The weight obtained by Simulated Annealing Algorithm has greatly avoided the appearance of local optimal solution and has accelerated the learning pace.

5. Experiments and Results

5.1. Processing Procedures

First of all, for conveniently extracting timbre features from music, the existing 2000 piano monotone audio database are chosen as sample and converted into mono, being classified into training data and test data. With regard to the violin music with smaller sample size, a suitable sample is selected from the ensemble to construct and complete on the basis of using the above specified construction rules, being carried out identically. The timbre feature is extracted by Mel cepstrum coefficient, also being unfolded as a one-dimensional vector, which is used as the input of BP neural network. Among them, 1500 are randomly selected as training data respectively, with the remaining 500 as test data. To improve the learning speed and get over the problems like local optimal solution, the Simulated Annealing Algorithm is used when setting the initial weight of the network, so as to obtain the approximate optimal solution of the initial weight. Meanwhile, the Sigmoid function is added as the transfer function at each output layer of the network. In order to solve the influence of noise on the accuracy of the results in the section, a threshold has been set in the network and without regard to the timbre that haven’t exceeded the threshold, which has increased the robustness of the network and bettered the learning pace. The initial learning rate is set as 0.01.

5.2. Comparison of amending network parameters

In the course of the experiment, for sake of avoiding the local optimal solution of the network owing to the limitation of training samples, existing single tone music period have been randomly classified for many times, which have generated different training sets and test sets correspondingly. Different servers are used to train the network and detect its training results. Most of the data results are stable and highly accurate, besides, the recognition results are almost the same. However, there arises a large error in the results when using one of the data sets for training, indicating the emergence of local optimal solution. For making the network have better adaptability and avoid the emergence of the local optimal solution, this paper has adjusted the number of hidden layers and the learning rate of the network, which have avoided the emergence of the condition after having tried to increase the hidden layer or improved its learning rate. At the same time, we have also tried to add additional supervision units to avoid the emergence of the situation. But in view of the condition of same result, adding additional supervision units will take longer training time, so it is not adopted. The comparison of the effects of different parameters is shown in Table 1.

| Learning rate | Hidden supervised Training rate | Normal Accuracy | Local optimal solution |
|---------------|--------------------------------|----------------|------------------------|
| 0.01          | 30 No                           | 0.7621         | Yes                    |
5.3. Comparison of BP algorithm before and after optimization
To test the effectiveness of the optimization method, in the course of the experiment, the contrast is conducted between networks the initial value not set by the Simulated Annealing Algorithm and not using reasonable sample space. The results show that the unoptimized BP network has slow learning rate and low accuracy rate, what’s more, it is prone to the phenomenon of local optimal solution. The results of the experiment are shown in Table 2.

| Simulated annealing algorithm | Planning sample space | Training rate | Accuracy  | Local optimal solution |
|------------------------------|-----------------------|---------------|-----------|------------------------|
| No                           | No                    | Very slow     | 0.7023    | Yes                    |
| Yes                          | No                    | normal        | 0.8165    | No                     |
| No                           | Yes                   | slow          | 0.8917    | No                     |
| Yes                          | Yes                   | fast          | 0.9105    | No                     |

5.4. Comparison with other feature extraction methods
In order to produce a more effective and intuitive comparison with other methods, several methods commonly used in audio feature extraction are selected for control experiments. The results show that these methods have two kinds of problems, the first is they can’t accurately identify multiple timbre, and the next is the recognition range is narrow when identifying a single timbre. There is also a certain gap in the recognition accuracy compared with the method in the paper. The robustness of the network is not ideal when adding artificial disturbance. The results of the experiment are shown in Table 3.

| Feature | Recognition rate of Multi-timbre | Recognition rate of Single-timbre | Recognition range |
|---------|----------------------------------|----------------------------------|-------------------|
| CQT     | 0.9172                           | 0.9423                           | Narrow            |
| NMF     | 0.7314                           | 0.9145                           | Normal            |
| STFT    | 0.8936                           | 0.9172                           | Very narrow       |
| PLCA    | 0.7025                           | 0.8976                           | normal            |

5.5. Results of the experiment
In this experiment, four evaluation criteria, accuracy (Accuracy), precision (Precision), recall rate (Recall) and comprehensive index (F-Score) are adopted to evaluate and analyze the experimental results. Due to the randomness of the network, several comparative experiments are carried out under the same training methods and different training samples, so as to make the experimental results more convincing. The experimental results are shown in Table 4.

| Number of inputs | Accuracy | Precision | Recall | F-Score |
|------------------|----------|-----------|--------|---------|


6. Conclusion

In this paper, the existing single tone music period is used as training data of neural network, and MFCC is selected to extract timbre features, input into the BP neural network for training and recognition, which have finally reached high accuracy and stability. In the process of constructing network and training samples, the problem caused by insufficient sample size is resolved by making reasonable screening rules and planning the sample space. The Simulated Annealing Algorithm is utilized to find the approximate optimal solution in determining the initial weight of BP neural network, which has greatly shortened the training time, avoided the emergence of local optimal solution, and improved the accuracy. In the processing of future research, we should take further trial in the impact of different network layers and learning speed, and simultaneously add distinctive timbre to optimize the network. At the same time, we should also try to collect audio data independently for experiments.

References

[1] Zeng, C. (2019). Research and Implementation of Key Technologies foe Automatic Transcriptio n of Multi-Instrument Music. https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD201902&filename=1019113063.nh.
[2] Xv, Y.Z. (2018). Research and Application of Music Classification Based on Convolutional Neural Network. https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD201901&filename=1018898729.nh
[3] Zhang, F., Li, L. (2019) Semi-supervised Learning Approach Based on Non-negative Matrix Factorization and Harmonic Functions. Statistics & Decision., 22:71-73.
[4] Siddharth, S., Emmanouil, B., Simon, D. (2015) An end-to-end neural network for polyphonic piano music transcription. https://arxiv.org/abs/1508.01774.
[5] Wu, J.J., Meng, L.L. (2009) The recognition of the musical pitch. Electronic Measuremen-t Technology., 32(4):126-128.
[6] Xu, P.J., Guo, L., Liu, S.C. (2011) Note recognition combining pitch-detection and onset-detecti on. Journal of Computer Applications., 31(s2):172-175.
[7] Chen, Y.W., Li, K., Han, Y., Wang, Y.P. (2019) Musical Note Recognition of Musical Inst ruments Based on MFCC and Constant Q Transform. http://kns.cnki.net/kcms/detail/50.1075TP.20191122.1411.008.html.
[8] Springer, D., Gregory, A.L., Lewis A.J., et al. (2021) Effects of Dark and Bright Timbral Instructions on the Production of Pitch and Timbre. Journal of Research in Music Education., 68(4):482-498.
[9] Li, Z.P., Zhao, D.Q., Xiang, M.Z., Zhang, L.W. (2019) A Short-Time Fourier Transform Algorithm for Accelerating GNSS Signal Acquisition. Journal of Geomatics Science and Technology., 36(01):23-27.
[10] Todisco, M., Delgado, H., Evans, N. (2017) Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. Computer Speech & Language., 45:516-535.
[11] He, L., Yuan, B. (2019) Classification of Music Genres Using Long and Short-term Memory Network. Computer Technology and Development., 29(11):190-194.
[12] Lin, L., Wang, R.D., Yan, D.Q., Li, C. (2018) A playback speech detection algorithm based on log inverse Mel-frequency spectral coefficient. Telecommunications Science., 5:91-98.
[13] Yang, Y., Chen, J.J. (2020) Review on Application of Intelligent Algorithm to Optimize BP Neural Network. Computer Knowledge and Technology., 16(35):7-10+14.

[14] Qin, L., Pan, Y.R. Endowment insurance subsidy intervention risk assessment system based on improved BP neural network. Modern Electronics Technique., 43(24):156-159.

[15] Li, Y.N., Zeng, Q.H., Zhang, Y.Y., Jiang, Y., Cui, Y.C. (2020) Mura detection and positioning in picture based on BP neural network. Opto-Electronic Engineering., 47(11):63-70.

[16] Luo, G., Qi, Y.P., Liang, S., Luo, J.Y., Jia, C.L. (2020) Safety Assessment of The Ancient Timber Frames based on BP Algorithm. In: Industrial Building Academic Exchange Conference 2020. Beijing. pp. 1683-1686+1749.

[17] Lin, B.L., Zhao, Y.N., Lin, R.X., Liu, C. (2021) Integrating traffic routing optimization and train formation plan using simulated annealing algorithm. Applied Mathematical Modelling., 93:811-830.

[18] Yang, Y., Li, K., Wang Y.G. (2020) Application of BP neural network based on simulated annealing algorithm in bridge intelligent maintenance. Intelligent City., 6(21):12-13.