IOP Conf. Series: Materials Science and Engineering **435** (2018) 012005 doi:10.1088/1757-899X/435/1/012005

Singer Recognition Based on Convolutional Deep Belief Networks

Yang Li[1] and Chu Li[2]

[1]School of Information Technology, Shanghai Jianqiao University, Shanghai 201306, China
[2]Pudong Foreign Languages School, Shanghai 201203, China

E-mail: 16083@gench.edu.cn

Abstract. Singer recognition is an important branch of music retrieval and classification. This paper focuses on the application of convolutional deep belief networks (CDBN) for singer recognition. First, the system architecture of singer recognition based on CDBN is given, and then the pre-processing of the song signal is described in detail, including sampling, framing, pre-emphasis and windowing. The feature extraction of song signal based on MFCC is described in detail, and the composition and principle of CDBN and its application in singer recognition are introduced. Experiment based on three different feature extraction techniques of LPCC, MFCC and CDBN is carried out and the result is compared and analysed, the experimental results show that CDBN is effective for singer recognition.

1. Introduction

With the rapid development of computer technology, network technology and multimedia technology, the amount of audio data on the Internet is increasing rapidly. How to effectively retrieve and classify audio information is a very important research topic [1].

Singer recognition is an important branch of music retrieval and classification. Its main purpose is to distinguish the singers from the unknown songs and find out which singers they belong to, so that they can automatically classify the songs by different singers [2]. Singer recognition can also be used for copyright protection, and record companies can automatically scan whether the songs provided by suspicious websites contain music that belongs to the company's copyright. In addition, self-service singing operators can use singer recognition technology to get users' favourite music types, so as to recommend some personalized music services.

Deep learning is a deep simulation of the human brain, and it can acquire the characteristics needed by itself. The deep network can analyse the input data by itself, and learn the characteristics of the data classification, which reduces the workload of the artificial extraction of classification features, and makes the classification recognition system more intelligent [3]. Therefore, it is very meaningful and valuable to study deep learning framework model and apply deep learning knowledge to music information retrieval task. This paper focuses on the application of convolutional deep belief networks (CDBN) in singer recognition.

2. Singer recognition system architecture

The singer recognition system is basically divided into two parts: training and testing. As shown in Figure 1, the training stage is mainly composed of selected training samples, pre-processing, feature

extraction and CDBN training & modelling. The test stage is mainly composed of test samples input, pre-processing, feature extraction and CDBN testing [4]. Among them, the same pre-processing part of the two stages includes framing, windowing and pre emphasis. In the pre emphasis stage, the frequency spectrum of the signal can be easily analysed and processed by enhancing the high frequency signal in the signal. In the framing and windowing parts, the song signal is divided into a series of song frames based on Hamming windows based on the short time stationary characteristics of song signals. The feature extraction stage is the analysis and extraction of the pre-processed song frames in the time domain and frequency domain, so as to describe the characteristics of these song frames. For example, Mel frequency cepstrumcoefficient (MFCC) and its variants. The extracted feature vectors are compared with each singer's CDBN model to determine the result of recognition.

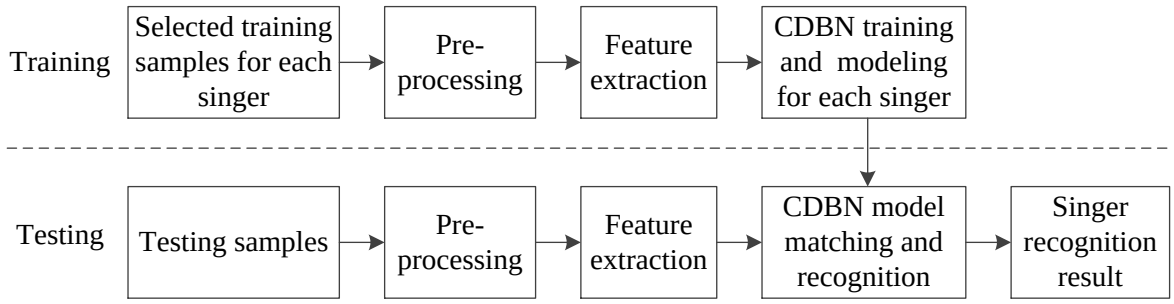

Figure 1. Singer recognition system architecture.

3. Pre-processing of song signals

3.1. Sampling

Both songs on the Internet and songs in CD belong to digital music. They all have a certain sampling frequency and encoding mode. The way of music coding and the frequency of sampling have a great influence on the analysis of music information. Therefore, it is necessary to unify the sampling frequency and coding mode of the music signal before framing. In most applications, digital music does not require too high sampling rate, and the high sampling rate will increase the computational complexity. In this paper, the sampling frequency of 22.05kHz is adopted, thus the efficiency of the algorithm is improved. Under the premise of not losing the basic recognition characteristics of music, the test music signal (44.1kHz) can be sampled, which has little influence on the recognition effect.

The song signals that are downloaded from the Internet can not be directly used to extract feature vectors. Before analysing and extracting the features, they should be handled in a unified way to meet the requirements of extracting features. This section introduces the pre-processing method. This stage mainly includes four steps: sampling, framing, pre-emphasis and windowing.

3.2. Framing

Considering the time continuity and short-time stability of the song signal, song signals should be processed in frames after sampling. The length of the frames will directly affect the extraction and recognition results.In order not to lose the information of the song signal change, the sliding window is used to divide the frame, and the window function smoothly slides on the song signal, and the signal is divided into frames. The framing can be continuous or overlapping, so that there is an overlap between the frames. The overlapping part is called frame shift. The frame shift value is generally 1/2 of the frame length. The framing is usually used with the windowing process described in section 3.4.

3.3. Pre-emphasis

Because of the radiation at the lip, the energy loss of speech or song signal energy is obviously higher than that at high frequency. Therefore, the aim of pre-emphasis is to improve the high frequency part

of the speech signal, so that the frequency of the high frequency part becomes gentle, which helps to analyse the spectrum of the whole speech signal.

The influence of the energy loss of the speech or song signal is shown in the relationship between the frequency and the power spectrum: if the frequency of the speech signal is increased by two times, the amplitude of the power spectrum will be reduced by about 6dB. Therefore, theoretically, the pre-emphasis of the high frequency part should also be processed on the 6dB/oct value. In this way, the amplitude of the high-frequency part of the signal is increased by pre-emphasis, and the amplitude is not much different from that of the middle and low frequency parts.

Pre-emphasis can be implemented by hardware or software. This paper adopts software method. The method is to use a digital filter to process the sampled signal. The transfer function of the filter is: $H(z)=1-az^{-1}$, among it, a the pre-emphasis factor, the value is usually close to 1, such as a=0.9375.

3.4. Windowing

Both above framing and pre- emphasis are inseparable from the speech signal "short time analysis technology". The so-called short time analysis technology is based on the theory of short time stationarity of speech signals. Although the speech signal is constantly changing with time, however, in a very short time range, it is generally considered that the short time is 10-30ms, the characteristics of the speech signal are basically unchanged, which can be regarded as a short and stationary procedure, that is, the short time analysis technique can be used to process the speech signal. Correspondingly, in the short period of 10-30ms, the model parameters representing the characteristics of speech signals will not change.

The purpose of windowing is to divide the song signal into frames, and the frame length is N. For each frame, the time window function $\omega(n)$ is multiplied with the original voice signal $s(n)$, which contains the speech signal sequence of N points in the sample, and N is the length of the window.

There are usually two kinds of window functions commonly used in windowing.

(1) Rectangle window function:

$$\omega(n)=\begin{cases}1, & 0\le n\le N-1\\ 0, & otherwise\end{cases} \tag{1}$$

(2) Hamming window function:

$$\omega_H(n)=\begin{cases}0.54-0.46cos\left(\frac{2\pi n}{N-1}\right), & 0\le n\le N-1\\ 1, & otherwise\end{cases} \tag{2}$$

From the two window function formula, we can see that the window length of the window function has a great impact on the signal. When the length of the window is large, the window function can be regarded as a narrow low pass filter, the length of which is close to the period of several fundamental tones. The short time information of the signal is almost same, the detail of the waveform can be ignored; the length of the signal may be close to or even less than the cycle of a fundamental tone when N is relatively small, and the change of short time energy of the signal is obvious, and can't get smooth short time information.

The rectangular window is usually used in time domain analysis, and Hamming window is usually used in frequency domain analysis. It can be seen from the formulas that the boundary of the rectangular window is discontinuous, and the leakage will appear in the calculation. But the function boundary of the Hamming window is smooth, which can effectively avoid the leakage phenomenon.

4. Feature extraction of song signals

Feature extraction is crucial for singer identification. the features can present human voice include: short time energy, short time average amplitude, short time pitch period, linear prediction coefficient (LPC), partial correlation coefficient (PARCOR), line spectrum pair (LSP), short time spectrum,

cepstrum feature, Mel frequency cepstrum coefficient (MFCC) and so on. MFCC is an auditory feature based on human ear, and it is also the most widely used feature parameter in the audio recognition model [5].

Human ears have different perceptual abilities to sound signals of different frequencies. In general, the human ear perception ability and the sound frequency are basically linear in the frequency region below 1kHz. But in the area above 1kHz, the perceptual ability of human ear is logarithmic with frequency. In order to simulate the special perceptual characteristics of human ears, the concept of Mel frequency was put forward. The conversion formula between the linear frequency of sound and Mel frequency is as follows:

$$f_{mel} = 2595 \times \log_{10}\left(1 + \frac{f}{700}\right) \tag{3}$$

The feature extraction process of MFCC is:

(1) Pre-processing of song signals, as described in section 3.

(2) First, fast Fourier transform (FFT) is used to get the spectrum of each frame, then the spectrum is squared to get the amplitude spectrum of each frame. The discrete Fourier transform (DFT) of the song signal is:

$$X_a(k) = \sum_{n=0}^{N-1} x(n) e^{-\frac{j2\pi kn}{N}}, 0 \le k < N \tag{4}$$

(3) A set of Mel filters is designed to filter the amplitude spectrum of the previous step. The logarithmic energy of the output of each filter group is calculated as:

$$\mathrm{s(m)} = \log_{10}\left(\sum_{k=0}^{N-1} |X_a(k)|^2 H_m(k)\right), 0 \le m \le M \tag{5}$$

(4) Then discrete cosine transform (DCT) to get MFCC:

$$\mathrm{C(n)} = \sum_{m=0}^{N-1} s(m) \cos\left(\frac{\pi n(m-0.5)}{M}\right), n = 1,2 \ \ldots, L \tag{6}$$

In this paper, we first pre-process the song signal and extract the MFCC feature, and then use this feature as the input of the deep belief network.

5. Singer recognition based on CDBN

5.1. The basic idea of deep learning

Deep learning is a new field in machine learning. It is similar to the neural network, which is similar to the layered structure of the biological brain. Its structure is that the system is a multi-layer network consisting of the input layer, the hidden layers (multi layers), the output layer, and each layer of the network can be regarded as a logistic regression model, the adjacent layer nodes of the model are interconnected, and there is no connection between the same layer and the cross layer node.

The basic idea of deep learning is for multiple stacking layers, the input of this layer is the output of the upper layer, that is, the hierarchical expression of input information. By adjusting the parameters in the system, a series of hierarchical features can be obtained automatically [6]. The process of extracting network parameters is widely used in feature extraction and pattern classification.

5.2. Restricted Boltzmann machine

Restricted Boltzmann machine (RBM) is energy based generative structural model, which is usually used to construct deep neural network structure [7]. RBM is an undirected graph, contains two layers of the visual layer v and the hidden layer h, the neuron nodes are not connected in the respective layers of the visual and hidden layers, but the interlayer is fully connected, as shown in Figure 2.

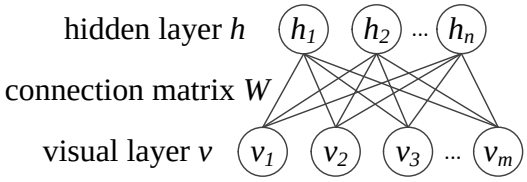

Figure 2. Restricted Boltzmann machine.

From Figure 2, we can see that there are *m* nodes in the visual layer of the RBM, which are used to represent the input data. The hidden layer has *n* nodes, which is used to extract features, the *W* representation layer to layer neuron connection weight matrix, and the calculation formula of the energy function is as follows:

$$E(v,h,\theta) = -\sum_{i=1}^{m}\sum_{j=1}^{n} v_i h_j w_{ij} - \sum_{i=1}^{m} b_i v_i - \sum_{j=1}^{n} a_j h_j \quad (7)$$

Among them, v_i and b_i represent the *i* unit of visual layer and its bias, h_j and a_j represent the *j* unit of hidden layer and its bias, and w_{ij} represents the weight of the connection between the *i* unit of visual layer and the *j* unit of hidden layer, θ is the actual parameter used to limit the RBM.

5.3. Deep belief network

In 2006, Hinton proposed the concept of deep belief network (DBN) and proposed an efficient learning algorithm [8]. It is a greedy layer by layer unsupervised training algorithm. It is a complex directed acyclic graph, consisted of RBM, so the same layer is not connected. The relationship between visual layer and hidden layer of DBN can be expressed by the joint probability distribution:

$$P(v,h^1,h^2,\ldots,h^l) = P(v|h^1)P(h^1|h^2)\ldots P(h^{l-2}|h^{l-1})P(h^{l-1}|h^l) \quad (8)$$

In this formula, *l* is the number of the DBN hidden layers.

DBN adopts the method of training RBM from low to high and layer by layer, and unlabelled data is used during this stage. After this pre-training, the traditional global learning algorithm, such as reverse propagation or wake-sleep algorithm, can be used to further adjust the network by using the labelled data, that is, to use the maximum like function as the target function to make the network reach the best. DBN overcomes the high complexity of DNN global training and converges to the local optimal solution caused by improper training of BP. The training mode of DBN can be summarized as two parts: first, the better initial parameter values are obtained by training RBM layer by layer, and then the network is further optimized through the traditional learning algorithm. Because the training method combining DBN and traditional learning algorithm has many advantages, it has been widely applied in the field of speech recognition.

5.4. Convolutional deep belief network

DBN has good flexibility, it can be easily extended to other networks or combined with other models, and one of the typical DBN extensions is the convolutional deep belief networks (CDBN). CDBN is a stack based on the convolution restricted Boltzmann machine (CRBM). CRBM is also divided into hidden layer and visual layer. Unlike RBM, the connection between the CRBM visual layer and the hidden layer is not a full connection. Each hidden layer node is connected only to a portion of the visual layer nodes, and each visual layer node is connected only to a portion of the hidden layer nodes. All hidden nodes are operated by the same convolution kernel. CRBM is characterized by partial receptive field and weight sharing, that is, the hidden layer and the visual layer are partially connected, and the weights of the models are shared. Weight sharing can reduce the number of parameters to be trained in the neural network. CDBN is a hierarchical probability generating model [9]. CDBN can be understood as a combination of DBN and convolutional neural network (CNN).

The training mode of CDBN is similar to DBN, unsupervised training is first carried out, and CDBN is trained using CRBM layer by layer. After pre-training, a supervised fine tune is adopted, and the weight value updating formula during the reverse propagation process is the same as in CNN [10].

The advantages of CBDN are:

(1) Absorbing the unsupervised training method of DBN, using a large number of unlabelled samples to initialize weights.
(2) Absorbing convolution concept of CNN, can use different convolution kernel to capture different features in music or song.
(3) By increasing the convolution step size, can reduce the number of hidden layer nodes.
(4) By sharing the weight, can reduce the number of parameters to be trained in the neural network.

CDBN is used in this paper for singer recognition because of above advantages.

6. Experiment and result analysis

6.1. Experimental environment

The experiment is done on personal computer. The software and hardware environment includes: CPU is Intel i7-4700, memory size is 16GB, operation system is Linux Ubuntu l6.04 LTS Server, deep learning framework is TensorFlow 1.0.1, programing software is Python2.7.

6.2. Song samples selection

In singer recognition experiment, the selection of singer has a great impact on the recognition result, because different singers may have different identities, and there are no open music databases for singer recognition. In this experiment, 10 singers were selected, with 5 male singers and 5 female singers. Each singer selected five songs as training samples, and the other five songs as test samples.

6.3. Pre-processing

Because songs are downloaded from the Internet, the encoding format is MP3 format, and the sampling rate is 44.1kHz, dual channel. In this paper, a unified sampling of these songs is used, the sampling rate is 22.05kHz. And cut the song into samples of 3s length. Because the song contains pure musical segments and silent segments, these segments have big impact on singer recognition. In this paper, samples of the 3s length segments are screened. If the singer's voice length is less than 1s, the sample is regarded as a pure musical segment and is discarded.

6.4. Feature extraction

For the CDBN, the number of convolution kernel used in this paper is 300, the size of convolution kernel is 8, the number of iterative rounds is 100, and the number of training times in each round is 100. The initial value of the connection weight used in the first tier model is the Gauss distribution random value with a mean value of 0 and a variance of 0.01. The bias of the visual layer takes is 0, and the bias of the hidden layer takes -0.1.

In order to compare the effect of feature extraction, the traditional features are also extracted. The traditional features mentioned here refer to LPCC feature and MFCC feature.

6.5. Comparison and analysis of experimental results

Because the focus of this experiment is to compare the feature extracted by deep learning with the traditional single feature, instead of comparing the model of the classifier, this paper uses the support vector machine (SVM) model as the classifier [11]. In the experiment, the three features and their corresponding classification labels are input into the SVM model respectively to train the classifier. After obtaining the classifier model, song segments of the 10 singers are classified and recognized. The recognition results are shown in Table 1.

Table 1. Recognition result comparison.

	LPCC	MFCC	CDBN
Recognition rate	53.6%	58.4%	77.8%

According to the results in Table 1, the recognition rate of the three features is not very high, mainly due to the interference of background music. But the recognition rate of CDBN is much higher than the other two features, that shows CDBN has better recognition effect on singer recognition than traditional LPCC and MFCC features. It proves that CDBN is effective in singer recognition.

7. Conclusions

On the basis of reading a large number of literatures about singer recognition, music information retrieval and speech processing, this paper studies the singer recognition from several aspects, such as the pre-processing of song signal, the extraction of auditory feature parameters, singing recognition and the establishment of the singer model. As a new feature extraction technology, deep learning has achieved success in the field of speech signal processing. This paper draws on the research results of deep learning in speech signal processing, based on the combination of music classification and deep learning theory, analyses the characteristics of CDBN in detail, and uses CDBN for singer recognition. In the experiment, for the music songs with background accompaniment on the Internet, the feature extraction based on LPCC, MFCC and CDBN are compared. The experimental results show that the CDBN learning features have better effects than traditional LPCC and MFCC features in singer recognition. The experimental results show that CDBN has great research value in singer recognition.

References

[1] Li T and Ogihara M 2006 Toward intelligent music information retrieval *J. IEEE Transactions on Multimedia***8(3)**564-574
[2] Yuan T T and Cao M M 2015 Voice-recognition-based music retrieval system *J. Bulletin of Science and Technology***31(7)** 156-159
[3] Liu F Y, Wang S H and Zhang Y D 2017 Survey on deep belief network model and its applications *J. Computer Engineering and Applications***54(1)** 11-18
[4] Gong A, Jing M B and Dou F 2017 Music mood classification method based on deep belief network and multi feature fusion *J. Computer Systems and Applications***26(9)** 158-164
[5] Lv L L 2016 Singing voice detection in songs based on clustering of MFCC *J. Computer Knowledge and Technology***12(31)** 170-171
[6] Xu X X 2017 Study on gesture recognition based on PCA and DBN *J. Artificial Intelligence***36(13)** 55-58
[7] Salakhutdinov R and Hinton G 2012 An efficient learning procedure for deep Boltzmann machines *J. Neural Computation***24(8)**1967-2006
[8] Hinton G E and Salakhutdinov R R 2006 Reducing the dimensionality of data with neural networks *J. Science***313(5786)** 504-507
[9] Yi L and Ya E 2017 Based on Gabor feature and deep belief network face recognition methods *J. Computer Simulation* **34(11)** 417-421
[10] Brosch T, Tam R 2015 Efficient training of convolutional deep belief networks in the frequency domain for application to high resolution 3D images *J. Neural Computation***27(1)** 211-227
[11] Chen W H 2016 Emotion classification of music based on support vector machine *J. Software Engineering***19(12)** 20-23