Research on music genre classification algorithm based on deep learning

Zhang Haoran

Music genre classification/recognition is an essential and critical task in the field of music information retrieval, and it is a necessary processing link in music recommendation and music automatic labeling. Music genre classification is very useful for users to search for their favorite music works. Music genre classification has always been a core problem in understanding human preferences for music, and has been applied to the construction of music recommendation systems. In this paper, long short-term memory (LSTM) model is used for music genre classification instead of convolutional neural network. By training a deep model, it is possible to classify ten different types of music. In addition, this paper also adopts a hierarchical classification scheme to further improve the accuracy. First, by training the LSTM classifier, music is divided into two classes: fortissimo class and weak tone class, and then music is further divided into multiple subclasses.

***Index Terms*—Music information retrieval; Music genre classification; LSTM.**

I. Introduction

At the beginning of the 21st century, the world is facing the continuous growth of online music information, and the penetration of the Internet into daily life has enhanced the development of online music information. Efficient and accurate automatic music information processing (especially access and retrieval) will be an extremely important topic, which has been receiving more and more attention. Music can be classified according to its genre, which has a hierarchical structure.Music genres in online databases are one of the main structures for organizing huge amounts of music data. Music genres are designated by human experts and amateurs (such as users), and the process of labeling is time-consuming and expensive. Because it is extremely difficult to give a precise definition of musical genres, the current genre classification is largely done by hand, and many musical sounds are ambiguous on the boundaries between genres. These difficulties are due to the fact that music is an evolving art, and performers and composers are influenced by other genres of music. However, it has been noted that audio signals (digital or analog) belonging to the same genre have certain characteristics in common because they are composed of similar types of instruments, have similar rhythmic patterns, and have similar pitch distributions, suggesting that automatic genre classification is feasible.

In terms of the time and manpower required to organize a music collection, manual classification of music genres became very expensive due to the need for more efficient tools to browse, organize, and dynamically update the collection, so automatic classification of music genres was required [4]. When automatic music genre classification is available, it has a positive effect on the generation of music recommendation system and automatic playlist. These belong to the broad field of music information retrieval. Music information retrieval plays a key role in digital audio processing, search and retrieval.

Automatic music genre classification, in which the problem can be strictly defined, that is, the classification of musical signals into a unique genre through a computational analysis based on musical feature representation. Music genre automatic classification is a basic component of music information retrieval system. In this paper, the genre classification process in music is divided into two steps: feature extraction and multi-class classification. In the feature extraction step, information representing music is extracted from music signals. Feature extraction should be comprehensive (able to describe the characteristics of the music very well), compact (need to reduce the data storage) and efficient (do not require too much extraction computation). In order to meet the first requirement, it must be designed to include both low-level and high-level information about the music. In the second step, a mechanism (algorithm or mathematical model) is established for identifying the musical genre labels they correspond to from the representations of musical sounds based on their characteristics.

II. Related Works

This section summarizes the relevant work in the field of music information retrieval, with emphasis on the automatic classification system of music genres. Due to the increase in storage capacity and the widespread use of digital audio music, it is now easier to collect large amounts of music data from the Internet. Music information such as genre, emotion, and style can be used to manage metadata for large music libraries [1]. Aucouturier and Pachet[2] argue that musical genres can provide useful descriptions of musical content. Certain properties (such as timbre, beat, rhythm, and performance instruments, etc.) may affect the style of music. However, people often view music as a whole, rather than individually identifying sources, attributes, or isolated fragments. [3] Thus, there is a gap between the units used for human perception and those used for computer analysis. In addition, building consistent and resizable corpora for classification tasks remains a difficult task, as it is difficult to give precise definitions of each type [4]. Therefore, automatic genre classification remains a challenging task.

The use of machine learning-based models to automatically label music genres has opened up new possibilities in this area of research, with promising results. Feature extraction is a widely used and important method for predicting the actual label of any audio file. Many models such as deep neural networks (DNN)[5] combined with other models such as decision trees [6] and maximum posterior probability models [7] have also been applied to music genre classification. DNNS have also proven to be very good at processing and training large amounts of data. Similarly, other neural networks such as CNN use spectrogram features to predict the class of a given audio input file, and its prediction results have been very close to the correct fact. It is known that there are multiple genres and sub-genres in music theory, but it is difficult to accurately draw the boundaries between them. Therefore, an alternative approach should be further adopted to parse a large number of audio files into their respective types.

So far, most classification methods have been supervised learning with some unsupervised samples [8]. Supervised learning requires manual labeling, which is time-consuming, while unsupervised learning has poor performance and poor accuracy. Some popular supervised classification algorithms are support vector machines (SVM), nearest neighbor (NN), Gaussian mixture model (GMM), and linear discriminant Analysis (LDA). Recently, semi-supervised classification methods have become increasingly popular in the machine learning community because it is known to provide good accuracy while requiring very little labeling work. The field itself is relatively new, with little work on automatic classification of musical genres based on semi-supervised learning. The main work in this regard is Song et al., who also proposed a content-based classification of music genres, combining different musical characteristics to carry out feature fusion [9].

Recurrent neural network (RNN), as a deep learning model, has been widely used in sequence data and has the ability to learn temporal relationships [10]. Because of gradient disappearance and gradient explosion, it is difficult to learn temporal relationships over long periods of time. The emergence of long short-term memory (LSTM) and threshold regression unit (GRU) as variants of RNN can solve these problems. This article uses LSTM to handle the music genre classification task. LSTM can better learn long-term dependencies, which is more suitable for music signal processing problems. By adjusting the time-based gradient backpropagation, the problems of gradient disappearance and gradient explosion can be solved.

III. Problem Statement

Feature extraction plays a crucial role in any classification system, and several different features are often used in music genre classification to design descriptive features for specific tasks of pattern recognition systems. For audio signal classification, the characteristics belonging to music include key dimensions such as timbre, harmony, spatial position, melody and rhythm. After feature extraction, standard classifiers can be used for classification.

The spectrum envelope obtained from the fast Fourier transform of the signal can be used to classify the amplitude characteristics of the genre. The exploration of amplitude allows this paper to identify signal variations, noise, loudness, and many other spectral features that describe aspects of discrete-time signals used for automatic music genre classification. Frequency spectrum features are usually obtained

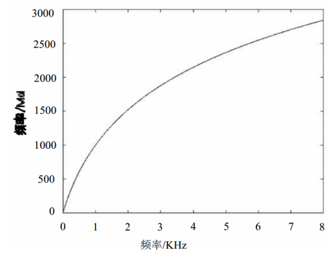

Fig. 1. Mapping between MEL frequency and actual frequency

by short-time Fourier transform. However, human auditory perception is not linear, as shown in Figure 3-1 is the mapping relationship between Meir frequency and real frequency. In fact, human auditory perception is in line with Meir frequency perception. Therefore, in this paper, Mayer frequency conversion is used to obtain the Mayer spectrum when using the spectrum characteristics. Timbre texture features are used to distinguish mixed signals of sounds that may have the same or similar rhythmic and pitch content. The use of these features stems from speech recognition, where in order to extract timbre features, the sound signal is first divided into statistically stationary frames, usually by windowed operations at fixed time intervals. A window function (usually a hamming window) can eliminate the wedge effect and then calculate the time structure features of each frame and calculate the statistics for these features (such as mean and variance).

Spectral variability is used to measure the dispersion of data variability, i.e. the extent to which signals converge or spread. This can be achieved by measuring the standard deviation of the signal amplitude spectrum. Mel frequency cepstrum coefficients (MFCC) are the coefficients that together form the Mel frequency cepstrum.

Most music retains regular rhythmic forms, creating the impression of rhythm. In order to understand the nature of music for genre classification, rhythm must be understood and preserved as a characteristic description. In this section, this paper establishes a rhythm detection scheme for music genre classification. In this part, the beat histogram is presented as the key feature vector.

Energy is a fundamental descriptive feature used in speech and audio processing. The energy is measured by calculating the sum of squares of the discrete-time signal. As shown in Figure 3-3, the energy distribution in different beats is different according to time segmentation. By calculating the energy amplitude in beats, we can get the characteristics of beats, check the arithmetic average of the first n Windows of the signal (for the experiment in this paper, n=100 is taken), and calculate the scores below the average. This paper can calculate the percentage of silence present in the signal as the fraction of low energy. A beat histogram is a range of signal strength that produces rhythmic intervals, which is achieved by measuring the energy of n continuous Windows and calculating the fast Fourier transform results, a type of feature that will produce very large matrices, so it can be introduced easily.

Rhythm is characterized by repeated patterns of sound and

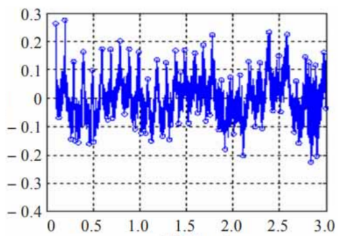

Fig. 2. Beat autocorrelation coefficient of Jazz

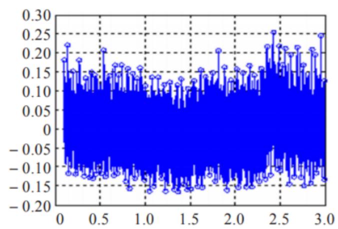

Fig. 3. Beat autocorrelation coefficient of rock music

the time of silence measured from the beat. It is this feature that can comprehensively describe the speed characteristics of music. Here is the related description of rhythm characteristics proposed by Pampalk in his paper [11].

Pitch content features describe the melodic and harmonic information of musical signals and can be extracted by various pitch detection techniques. The main peaks of the autocorrelation function are calculated, the sum of the envelope of each frequency band obtained by the decomposition of the signal is accumulated into the pitch histogram, and then the pitch features are extracted from the pitch histogram. The pitch content features usually include: the amplitude and period of the largest peak in the histogram, the interval between the two most prominent peaks, the sum of the histogram, etc.

The spectral feature extraction is introduced into the genre detection problem, which creatively combines the single char-

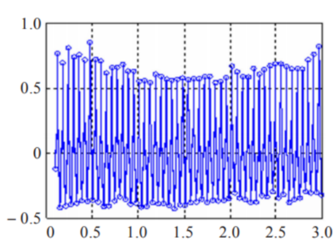

Fig. 4. Beat autocorrelation coefficient of funk music

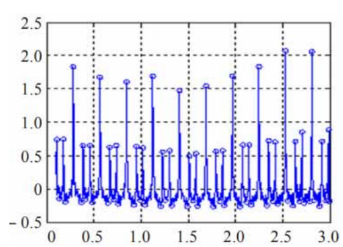

Fig. 5. Beat autocorrelation coefficient of electronic music

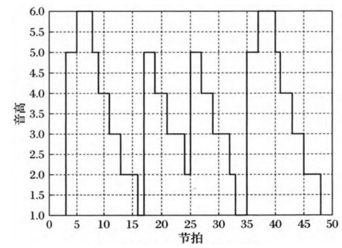

Fig. 6. Beat pitch visualization

acteristics of music with the chord structure and chord progression. Melodies and harmonies can be thought of as horizontal and vertical elements of music. Melody is identified by a series of note events, while harmonic is about the simultaneous occurrence of pitch and the use of different chords. Chroma[12] describes the melody and harmonic, and gives a representation of pitch information. It is a feature representation of a joint time dimension that gives intensity information related to each semitone within an octave. All octaves are represented as a square that describes the harmonic content of a piece of music. The most common method of Chroma feature extraction is based on Ellis, where the Chroma vector on the average beat length is represented by the Chroma vector on the beat synchronization. Since the time interval of different songs may be different, this paper selects only 30 Chroma vectors for each song randomly or sequentially, and connects them column by column, resulting in a 360-dimensional feature vector. Chroma has minimal influence on the playing instrument and rhythm, and the Chroma vector has information about the current note and is related to a particular chord.

No single feature vector can provide very good classification performance, because each music genre has different musical characteristics. Therefore, it is necessary to merge eigenvectors, and there are two possible ways to merge eigenvectors. The first method used in this paper is to weight the distance measure of each feature, while the second method can use a majority voting scheme when the classifier is output. This chapter introduces the feature set of feature extraction in the proposed music genre classification system. These features include amplitude-based features, which present timbre features describing loudness, noise, compactness, etc. Based on the characteristics of speed, it provides a method to explore the rhythm of signals. Based on pitch characteristics, an algorithm for describing the tone of a musical signal is presented; Finally, based on the characteristics of chord progression, this paper explores Chroma as a characteristic of chord (environment).

IV. Algorithms/Solutions

The Long Short-term Memory Model (LSTM) is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. Unlike standard feedforward neural networks, LSTMS have feedback connections that allow them to process not only individual data points (such as images), but also entire data sequences (such as voice or video). For example, LSTM is suitable for tasks such as undivided, connected handwriting recognition or speech recognition. A

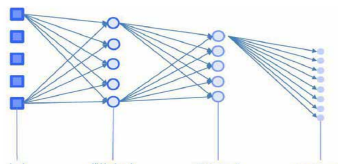

Fig. 7. Deep LSTM network architecture for music genre classification

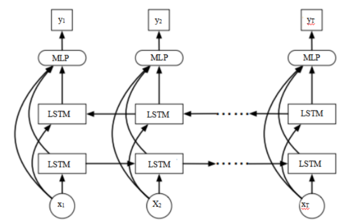

Fig. 8. Multi-layer LSTM classification network topology

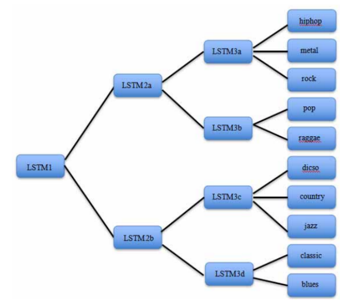

Fig. 9. The hierarchy of the LSTMs in the multi-step classifier.

typical LSTM unit consists of a memory unit, an input gate, an output gate, and a forget gate. The memory unit remembers values at any time interval, and three gates regulate the flow of information in and out of the memory unit. LSTM networks are well suited for classification, processing, and prediction based on time series data because there may be unknown duration intervals between significant events in a time series. LSTM was developed to solve the gradient explosion and vanishing problems that may be encountered in traditional RNN training, and the relative insensitivity to the length of the time interval is the advantage of LSTM over RNNS, hidden Markov models, and other sequence learning methods.

The deep LSTM network architecture used in this paper is shown in Figure 7. A four-layer network architecture is adopted in this paper. Among them, the first layer is the input layer, which contains 13 neuron nodes; the middle two hidden layers have 128 and 32 nodes respectively; the last layer is the output layer, which contains 10 nodes, and the distribution corresponds to ten music genre categories.

The LSTM network is trained using the training set data. The large amount of data is divided into ten genre labels. In the network training, there are 420 tracks in the data set for training, 120 for verification, and 60 for testing. Each track lasts 30 seconds. In this paper, the number of samples sent through the network for training is set at 35. Accuracy and loss are still improving in 20 years. At 20, the test accuracy is maximized and the loss is minimized. The classification accuracy of this paper is about 0.5 to 0.6, and there is still some room for improvement. With more training samples, this paper may be able to achieve an accuracy of 0.6 to 0.7, the main limitation being that the labeled training data is too small, which leads to low accuracy and overfitting. While some genres such as metal stand out and are easy to identify, it is also difficult to categorize some others that are quite similar. The classifier of LSTM architecture in this chapter is designed and proposed for music genre recognition. As shown in Figure 8, the details of the proposed network structure are: 2 recurrent neural network layers, and an RNN layer with gated recurrent units for time relation learning of the learned features. It is assumed that RNNS are best suited for time series aggregation features. After regularization, with the first layer (63x3x3) input of size (1x96x1400) as the size, the extracted features are pooled together by the maximum pooling of size (2x2), which results in the time dimension being reduced to 720 and the frequency to 48. So enter the second hidden layer number to become (63x48x720). A second layer with size (122x3x3), followed by a maximum pool size (3x3), gives the output size (8x16x240). The size of the output from the third layer (128x4x60) serves as the size of the input from the fourth and final layer (128x3x3). These pooling operations provide an output size of (128x1x15). This output is further processed to make it compatible with RNN networks. When looking at audio files of different genres, it is found that there are similarities between waveforms of different genres. Among them, the spectrograms of hip hop, metal, pop, rock and reggae swing have similar low frequency energy and high frequency energy from the image, while the spectrograms of jazz, disco, country, classical and blues are more similar, and a divide-and-conquer scheme is considered by further comparison. In the plan for this article, seven LSTM networks are used. Then, the 7 LSTM classifiers are used to design a multi-step hierarchical classification of 10 genres. The sample division for training and testing is the same as in Section 4.2. The LSTM classifiers involved are listed below.

LSTM1: It categorizes music into two categories: hard (hip hop, metal, pop, rock and reggae swing) and light (jazz, disco, country, classical and blues).

LSTM2a: It divides music into sub-forte 1(hip hop, metal and rock) and sub-forte 2 (pop and reggae swing).

LSTM2b: It divides music into two categories: secondary Mild 1(disco and country) and secondary Mild 2(jazz, classical and blues).

LSTM3a: It divides music into hip-hop, metal and rock. Similarly, training data were only sampled from hip-hop, metal and rock.

LSTM3b: It is used to distinguish between pop music and reggae swing.

LSTM3c: It is used to distinguish between disco music and country music.

LSTM3d: It is used to distinguish between jazz, classical and blues.

V. Evaluation

The GTZAN database is used for genre classification experiments and contains 1000 recordings covering 10 musical

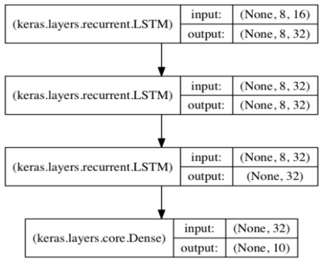

Fig. 10. LSTM-based music genre classifier network structure

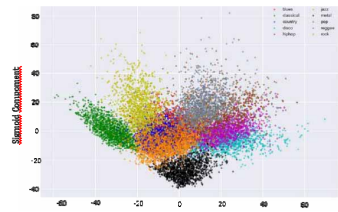

Fig. 11. Classification of music genres in GTZAN dataset

genres. Genres represented in the database include: Classical, country, disco, Hiphope, jazz, rock, blues, reggae, pop, and metal. Figure 5-1 shows the sample proportion of each category in the GTZAN dataset, where each type has 100 recordings. All recordings are mono, with a sampling rate of 22,050 Hz and a duration of approximately 30 seconds. Each recording is divided into 30 segments, so the duration of each segment is 1 second.

The GTZAN dataset is collected by Tzanetakis and is widely used for performance evaluation in the field of music. All stereo mp3 audio files are converted to mono waveform files with a sampling rate of 16khz during pre-processing. The GTZAN dataset contains ten music categories: blues, classical, country, disco, hip hop, jazz, metal, pop, reggae, and rock, each consisting of 100 30-second.au soundtracks. This paper randomly selects samples from the data set for training and testing. The test and training datasets do not overlap. The dataset was randomly split into 10 music folders, 9 of which were used for testing with the others. At the same time, the average accuracy of 10 fold cross verification is given to achieve the final test accuracy. dropout is set to 0.5 and learning rate is set to 1e-5. Under the condition that the convergence area of precision curve is stable, this paper stops training and saves the model.

In order to evaluate the proposed classification method and compare it with traditional genre classification methods, this paper performed a music genre classification task using a GTZAN dataset consisting of 1000 30-second music fragments with a sampling rate of 22050hz. Each clip is labeled with one of 10 genres, with 100 clips per genre.

The topology structure of the music genre classifier based on the LSTM model in the experiment is shown in Figure 10, which contains three LSTM layers. The input dimension of the first layer is consistent with the input feature vector dimension. The LSTM nodes of the middle layer employ 32 neurons. The final output layer uses a fully connected layer of 10 nodes for softmax classification to 10 different music genres. Due to the relatively small amount of data in the data set used in the experiment, the classification model proposed in this paper contains the deep learning model LSTM, which requires a large amount of data to train the model with strong learning ability, so it needs to be conducted before the experiment. In this experiment, the datasets used for training, validation, and testing the datasets were 8:1:1 and were distributed across different music genres. In order to accurately evaluate the effect of the model and the accuracy of the classification, the data set is divided into 10 parts in the experiment for cross-validation, which can ensure the accuracy and fairness of the model evaluation and avoid the influence of special case results on the accuracy of the model classification as much as possible. In the process of training the model, the optimization algorithm selects the learning rate adaptive Adam algorithm introduced above, and the learning rate decreases linearly from 0.002 to 0. The batch size selected from the training set for each training is 64, that is, the training set for each iteration is sampled, each training sample is trained once, that is, after a period of time, the samples in the data set are perturbed. During training, there will be fitting problems, which is often encountered when training deep learning models. Through continuous training of the model, the error on the training set will gradually decrease with the increase of the complexity of the model.

The two experimental techniques used in this experiment are described below:

EarlyStopping, training a deep learning model, is actually a process of constantly learning and updating parameters in the model through training data. In this process, with the deepening of the training process, the complexity of the model will gradually increase, the training error will gradually decrease, but the verification error will change from gradually decreasing to increasing again. This situation is called overfitting. Overfitting is a common problem in the training of neural networks. Early termination is a strategy to prevent overfitting. That is, parameters are set so that the model terminates the iteration before the iteration of the training set.

Early termination means that the accuracy of the validation set is calculated after one training session of all the training samples, i.e. the end of a period. The idea of early termination is that training stops if the accuracy of the model on the validation set does not continue to rise, but in the process of training the model, based on existing experience, it makes the accuracy on the validation set drop.

In this experiment, the six schools of similarity relationships are tested to see how the classifier distinguishes them. The configuration of the network nodes in this experiment is given in Table 5-1. Among them, 13 nodes in the input layer correspond to the input feature vector, that is, 13-dimensional MFCC features. 128 SIGmoids are selected as neurons in the first hidden layer, 32 SIGmoids are selected as neuron nodes in the second hidden layer, and 6 are selected as output nodes in the last layer.

The experimental results are shown in Figure 11. The standard neural network uses the Dropout neural network.

LSTM 分类器	准确度	次数
LSTM1	80.0%	35
LSTM2a	81.6%	20
LSTM2b	81.6%	35
LSTM3a	74.6%	40
LSTM3b	88.0%	20
LSTM3c	78.0%	20
LSTM3d	84.0%	40
最终	50.0%	N/A

Fig. 12. experiment time analyse

VI. Conclusion

With the popularity of Internet technology and multimedia devices, the amount of digital music available on all major platforms has increased dramatically. With such a large volume of music, it is difficult to manually manage and categorize the music. At the same time, users want to be able to quickly and accurately search the music they are interested in in the huge music database, which requires the design of an accurate and efficient system to manage a large number of music databases, and allow users to quickly and accurately search their favorite music types. In the field of music classification, the traditional classification method requires prior knowledge, and the process of feature extraction is complicated, so the features obtained are often lacking in generality. The classification model based on deep learning can realize automatic music classification by training the model.

Music signals can be thought of as time series data, and for this problem, classification models can be seen as coding-decoding frameworks. During the coding process, all input information should be compressed into a fixed vector representation and some information will be lost. Music information retrieval can help understand the context in audio signals and classify them according to various characteristics such as type, instrument, mood, etc. This classification gives users a better degree of differentiation when choosing their favorite music. More accurate retrieval of the labels present in music will give more users more opportunities to intelligently choose their favorite music. At present, deep feature extraction and machine learning methods can provide pre-labeled retrieval mechanisms, through which almost all content information in music can be understood. From local feature to time feature, the feature synthesis method using deep network is an exploratory label classification system.

This topic focuses on music genre classification, from audio feature extraction, classifier training to the final music genre prediction design and implementation of a complete music genre automatic recognition system. In this paper, we use long short-term memory model instead of convolutional neural network to classify music genres. In this paper, by training a deep model, music can be classified into ten different musical genres. In addition, this paper also adopts a hierarchical classification scheme to further improve the accuracy. Firstly, this paper divides music into fortissimo class and weak tone class by LSTM classifier, and then divides music into multiple subclasses. This new multi-stage classification method is used to classify different musical genres in each stage. The experiment shows that the classification accuracy can be improved to a certain extent by layering multiple steps.

References

[1] Scaringella N, Zoia G, Mlynek D. Automatic genre classification of music content: a survey[J]. IEEE Signal Processing Magazine, 2006, 23(2): 133-141.

[2] Tzanetakis G, Cook P. Musical genre classification of audio signals[J]. IEEE Transactions on speech and audio processing, 2002, 10(5): 293-302.

[3] Lidy T, Rauber A. Evaluation of feature extractors and psycho-acoustic transformations for music genre classification[C]. ISMIR. 2005: 34-41.

[4] Li T, Ogihara M. Music genre classification with taxonomy[C]. Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. IEEE, 2005, 5: v/197-v/200 Vol. 5.

[5] Baniya, B.K., D. Ghimire, and J. Lee, Automatic music genre classification using timbral texture and rhythmic content features, in The IEEE 17th International Conference on Advanced Communication Technology : "Named Data Networking - A Future Internet Architecture": IEEE 17th International Conference on Advanced Communication Technology (ICACT 2015), July 1-3, 2015, PyeonhChang, Republic of Korea. 2015: PyeonhChang. p.434-443.

[6] Dieleman, S. and B. Schrauwen, End-to-end learning for music audio, in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4-9 May 2014, Florence, Italy. 2014: Florence(IT). p. 6964-6968.

[7] Bengolea, G., et al, Feature Analysis for Audio Classification, in Progress in pattern recognition, image analysis, computer vision, and applications: 19th Iberoamerican Congress, CIARP 2014, Puerto Vallarta, Mexico, November 2-5, 2014, Proceedings. 2014: Puerto Vallarta(MX). p. 239-246.

[8] Agera, N., S. Chapaneri, and D. Jayaswal, Exploring Textural Features for Automatic Music Genre Classification, in First International Conference on Computing, Communication, Control and Automation: First International Conference on Computing, Communication, Control and Automation (ICCUBEA 2015), 26-27 February 2015, Pune, India. 2015: Pune. p.822-826.

[9]] Markov, K. and T. Matsui, Music genre classification using Gaussian Process models, in 2013 IEEE international workshop on machine learning for signal processing: MLSP 2013, September 22-25 2013, Southampton, United Kingdom. 2013: Southampton(GB). p. 1-6.

[10] Xu, J.-P., et al, Speech emotion recognition based on feature selection and extreme learning machine decision tree [J]. Neurocomputing. 2018. 273(Jan.17): p. 271-280.

[11] J. Andén and S. Mallat, "Deep Scattering Spectrum," IEEE Transactions on Signal Processing, vol. 62, pp. 4114-4128, 2014.

[12] G. Song, Z. Wang, F. Han, S. Ding, and M. A. Iqbal, "Music AutoTagging Using Deep Recurrent Neural Networks," Neurocomputing, 2018.