Music Genre Classification with LSTM based on Time and Frequency Domain Features

Yinhui Yi
School of Advanced Technology
Xi'an Jiaotong-Liverpool University
Suzhou, China
Yinhui.Yi20@student.xjtlu.edu.cn

Xiaohui Zhu
School of Advanced Technology
Xi'an Jiaotong-Liverpool University
Suzhou, China
xiaohui.zhu@xjtlu.edu.cn

Yong Yue
School of Advanced Technology
Xi'an Jiaotong-Liverpool University
Suzhou, China
yong.yue@xjtlu.edu.cn

Wei Wang
College of Computer and Cyber Security
Hebei Normal University
Shijiazhuang, China
wangwei82@msn.com

***Abstract*—Deep features generated from deep learning models contain more information for music classification than short-term features. This paper uses a long-short term memory (LSTM) model to generate deep features and achieve music genre classification. Firstly, two short-term features of Zero crossing rate (ZCR) and mel-frequency spectral coefficients (MFCC) are extracted from music in digital form, which is a time-domain feature and frequency-domain feature, respectively. Then these two features are fed to LSTM to generate deep features. Finally, we use support vector machine (SVM) and k-nearest neighbors (KNN) respectively to classify the music genre based on these deep features. Experimental results show that using LSTM can significantly increase the accuracy of music genre classification.**

***Index Terms*—Music classification; LSTM; Deep features; ZCR; MFCC**

I. Introduction

Listening to music is a common recreational activity in people's daily life. With the development of communication and network storage technology, large amounts of music are accessible to the general public. To better cater to listeners' preference, a new trend of effectively and efficiently retrieving music according to personal music interests is generated. One solution is to classify and assign labels to music based on genres, mood, artist, etc. [1]. In the early stage, the work of music classification is done manually. However, manual classification can be wrong and cannot provide large scale recommendation [2]. Now music companies are trying to solve this problem using efficient machine learning models. On the one hand, it can reduce the cost of managing and maintaining digital music collections because thousands of music needs to be uploaded to music libraries every year. On the other hand, the classification model can help users find music they are interested in and provide effective recommendation service.

The machine learning models for the music classification are generally made up of four steps: preprocessing, windowing, feature extraction and classification [3]. The most critical steps are feature extraction and classification. Fu et al. used hierarchical taxonomy to characterize audio features from different perspectives and levels as Fig. 1 shows [2].

The top level of Fig. 1 consists of semantic labels which can provide information relating to how people understand music. However, they cannot be obtained from low-level and mid-level features directly due to the semantic gap. The gap should be bridged by music classification algorithms [2].

From the aspect of music understanding, the features are classified into low-level and mid-level features. Low-level features are further divided into short-term (timbre) and long-term (temporal) features according to feature windowing duration. Mid-level features are closely related to features perceived by human listeners, mainly including rhythm, pitch, and harmony [4], [5]. They are extracted from low-level features. Short-term features can be obtained from music signals directly from time-domain or frequency-domain. They depict the tonal quality of the sound. Mel Frequency Cepstral Coefficients (MFCC) and zero-crossing rate (ZCR) are common short-term features in frequency-domain and time-domain respectively [6], [7].

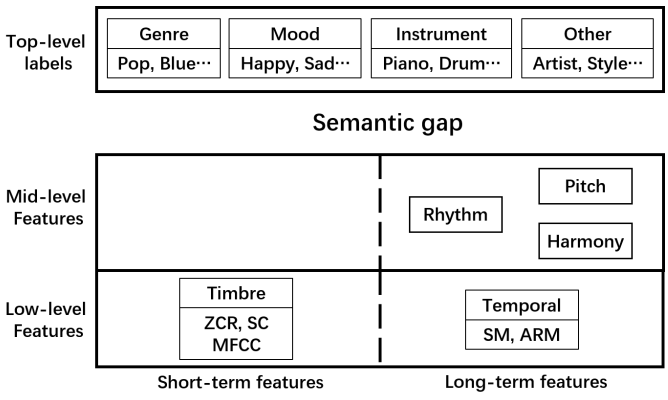

Fig. 1. Hierarchical taxonomy of audio features.

978-0-7381-2604-3/21/$31.00 ©2021 IEEE

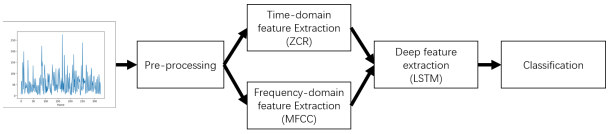

Fig. 2. Procedure of music genre classification using LSTM.

Long-term features depict the variation and evolution of short-term features over time. The simplest long-term features are statistical moments (SM) such as variance and mean from short-term features [2]. Additionally, long-term features can be extracted by the auto regressive model (ARM) [8]. In recent years, deep learning models have also been applied in feature extraction and proven powerful. Choi et al. proposed a convolutional recurrent neural network (CRNN) for music tagging and found it works better than convolutional neural networks (CNNs) [9]. Li used a bottleneck layer of deep neural network (DNN) to generate features and classified them with Bidirectional Long Short Term Memory (BLSTM) network [10]. Costa even extracted features from spectrograms obtained from audio signal with CNN, while the others mentioned models extracted features by processing digital data [11]. Features extracted from the deep neural networks are known as deep features, which contain high dimensional information. [3].

In this paper, the long-short term memory (LSTM) is used as the deep learning model to extract deep features from ZCR and MFCC. Two classification algorithms of support vector machine (SVM) and k-nearest neighbors (KNN) are used separately to classify deep features and evaluate the accuracy of music genre classification.

The rest of the paper is organized as following structure. Section II introduces the methodology used in this paper, including the procedure of music genre classification, extracting ZCR and MFCC and the architecture of LSTM. Section III describes details of the experiments and analyzes the results. We conclude the work in Section IV.

II. Methodology

A. Main Procedure of the Algorithm

The main procedure of music genre classification in this paper is shown in Fig. 2. It consists of four stages: preprocessing, low-level short-term feature extraction, deep feature extraction and classification.

Pre-processing is done on the audio signal, which involves noise filtering, silence reduction, normalization, etc. [3]. Additionally, to save computation cost, music summarization methods like k-means algorithm can also be employed [13]. At low-level short-term feature extraction step, ZCR and MFCC are extracted separately. ZCR is a widely used time-domain feature, while MFCC is a widely used frequency-domain feature. After that, both of them are fed to LSTM in deep feature extraction step. There are three kinds of output: (1) deep feature generated from ZCR; (2) deep feature generated from MFCC; (3) deep feature generated from the combination of ZCR and MFCC. In the last step, three kinds of deep features are classified by SVM and KNN separately.

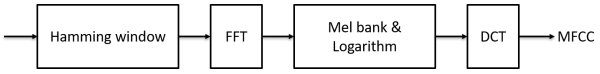

Fig. 3. Procedure of MFCC extraction.

B. Zero Crossing Rate

ZCR provides an impression regarding the frequency content. It depicts the rate of signal change in a time slot. A signal change is counted when signal changes from negative to positive or conversely [14]. ZCR at time T is computed as follow equation:

$$ZCR(T) = \frac{\sum_{n=1}^{L}[signal_n * signal_{n+1} < 0]}{L} \quad (1)$$

where L is the total number of signals in time slot T. If n^{th} signal, $signal_n$, times its next signal, $signal_{n+1}$, is negative, one change will be counted. ZCR is the sum of changes dividing the total number of signals L.

C. Mel Frequency Cepstral Coefficients

MFCC can convey the general frequency characteristics, which is close to the human audio system [15]. It is also extracted from fixed time slots. However, different from time-domain features, MFCC is in higher dimension. The procedure of extracting MFCC is illustrated in Fig. 3.

Music in waveform is firstly split into frames using Hamming window. Then for each frame, the Fast Fourier Transform (FFT) is employed to calculate the power spectrum. After that, a bank of triangular filters is applied to smooth the periodogram of the power spectrum. The center of frequencies of these triangular filters is spaced on the Mel-scale. Finally, take Discrete Cosine Transform (DCT) of the logarithm of filter-bank. The output is MFCC feature [10].

D. Long Short Term Memory

The basic unit of long short-term memory (LSTM) network is a memory cell [15]. Each memory cell t will process time-domain feature and frequency-domain feature at corresponding time slot t and generate new feature h_t as shown in Fig. 4. Each memory cell t is fed with the output from lower layer, x_t , and the information from previous cell, $cell_{t-1}$.

The internal structure of each memory cell is shown in Fig.5. It is a typical structure of LSTM containing forget gate, input gate and output gate [17]. One difference is that our structure contains two inputs of time-domain feature(ZCR) and frequency-domain feature(MFCC). They will be weighted by the model automatically.

The advantage of LSTM is that, on the one hand, music features are time-serial, which are suitable for LSTM. On the other hand, LSTM was designed to mitigate vanishing gradient problem of recurrent neural networks (RNN). Memory cells

of LSTM enable it to store and access information over long periods and generate better long-term features [12].

III. Experiments and Analysis

A. Dataset

The data set used in this paper is from Kaggle [18]. It is a music collection consists of 10 music genres with 100 audio files for each genre. The 10 music genres in the collection are blues, classical, country, disco, hiphop, jazz, metal, pop, reggae and rock. The audio files are saved in waveform with a length of 30 seconds. The property of each audio file is shown in table I.

B. Experiments

The raw dataset is a music collection saved in waveform format. Firstly, music is transformed into digital form. Each music in digital form is a one-dimensional array with a length about 660000 frames. Then ZCR and MFCC are extracted from the digital music data. The output ZCR feature set in time-domain feature extraction step is a 30×1 array for each piece of music, which means that music is sliced into 30 time slots and calculated ZCR for each. The output MFCC feature set is a two-dimensional array with 30×1293. This is not suitable to be the input of LSTM. Therefore, the mean and variance of each row were calculated as features. The final shape of MFCC feature is 60×1. In deep feature extraction step, the input shape of the LSTM layer is 700×30, 700×60 or 700×90 where 700 is the amount of music (only 70% is used to train the model). The number of input columns depends on the size of the feature set for each piece of music. The output of the LSTM layer is 700×128. 128 is the number of units in LSTM layer. The reason to use 128 units is that, after testing, more units cannot improve classification efficiency and accuracy. Additionally, the activation functions used in this project are the hyperbolic tangent and sigmoid,

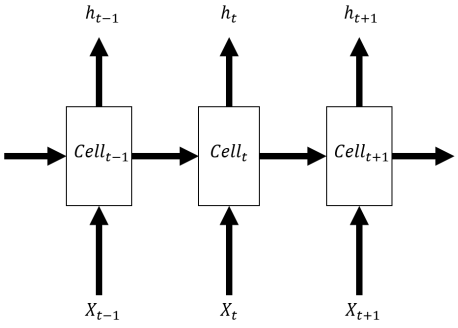

Fig. 4. Procedure of music genre classification using LSTM.

TABLE I
AUDIO FILE PROPERTY

Attribute	Value
Audio channel	1
Audio frames	Around 660000 frames
Sampling frequency	Around 22000Hz

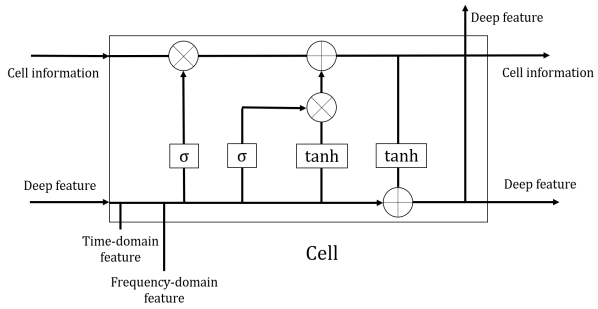

Fig. 5. Procedure of music genre classification using LSTM.

TABLE II
CLASSIFICATION ACCURACY WITHOUT LSTM

Short-term Feature	ZCR	MFCC	ZCR+MFCC
SVM	0.163	0.603	0.617
KNN	0.180	0.310	0.327

TABLE III
CLASSIFICATION ACCURACY WITH LSTM

Deep Feature	ZCR	MFCC	ZCR+MFCC
SVM	0.170	0.982	0.989
KNN	0.570	0.914	0.930

which are the default functions of the LSTM layer. Finally, the output of the LSTM will be fed to SVM and KNN. The trained classifier will be tested using the rest 30% of music. The average of 10 times execution results is used to evaluate the effect.

C. Results and Analysis

Table II shows the accuracy of music classification without deep feature extraction layer (LSTM layer). The lowest classification accuracy is only 0.163 when music is classified by SVM based on ZCR. While changing the classifier to KNN, the accuracy is still low. However, The classification accuracy significantly increases to 0.617 when music is classified by SVM based on the combination of ZCR and MFCC. Table II also shows that compared to KNN, SVM has better performance in all aspects.

Table III shows the classification accuracy using deep feature extraction layer. Compared to Table II, classification accuracy increases greatly in most cases regardless of SVM classifier based on ZCR. While classifier changes to KNN, the accuracy of the model based on ZCR increases by more than 200%, from 0.180 to 0.570. The highest classification accuracy appears on SVM classifier based on the combination of ZCR and MFCC with 0.989, which is slightly higher than only use MFCC.

Fig.6 illustrates the difference between using and not using LSTM. Deep features work well in most cases, which significantly increase the classification accuracy. However, when deep features are generated from ZCR and classified with

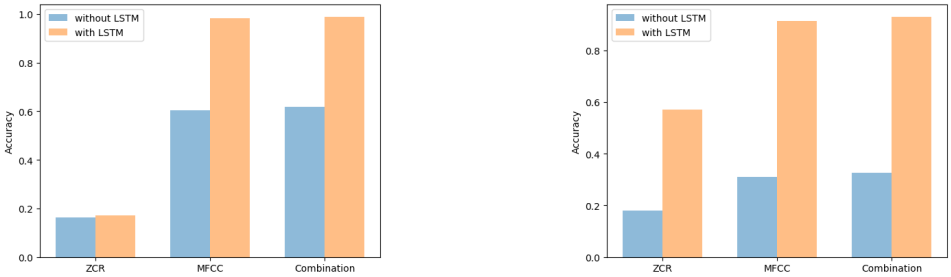

Fig. 6. SVM and KNN classification accuracy with and without LSTM.

SVM, accuracy does not improve obviously. This may result from SVM itself. When SVM classifies music, it tries to find hyperplanes that can make the distance from music vectors to hyperplane as large as possible. If the differences are large among music features, the effect of SVM is hard to improve. In contrast, KNN tries to cluster similar music. When features are amplified by LSTM, it can find more similar music easier. The advantage of MFCC is that it contains enough features initially. Thus, the differences among music are small. When the MFCC feature is further amplified with LSTM, the classification accuracy of SVM increases and is even better than KNN. However, when we combine the MFCC with ZCR , we find the classification accuracy increases slightly. We think that the quality and quantity of the feature set account for this. Only the MFCC feature and ZCR feature are used in this paper. Because the effect of MFCC is much better than ZCR, MFCC may dominate the classification. The improvement of feature combination based on MFCC and ZCR is limited. If more features can be fed to LSTM, the improvement may be more significant.

IV. Conclusion

In this paper, the effect of deep features generated from time-domain features and frequency-domain features for music genre classification is tested. Firstly, ZCR and MFCC are extracted for each piece of music which are in waveform initially. Then these features are fed to LSTM to generate deep features. Finally, deep features are classified by SVM and KNN respectively. The experimental results show that deep features generated by LSTM could improve the accuracy of music genre classification. In the future, we will continue to improve the algorithms including: (1) Test the current model with other music collections; (2) Use more kinds of low-level short-term features to feed the deep learning model; (3) Optimize the deep learning model by using bidirectional LSTM and Advanced LSTM; (4) Further explore the relationship between deep features and mid-level features.

Acknowledgment

This work was partly supported by the AI University Research Centre (AI-URC) through XJTLU Key Programme Special Fund (KSF-P-02 and KSF-A-19) and Research Development Fund of XJTLU (RDF-19-02-23).

References

[1] Z. Fu, G. Lu, K. M.Ting, and D. Zhang, “A Survey of Audio-Based Music Classification and Annotation,” in IEEE transactions on multimedia, vol. 13, no. 2, pp. 303-319, 2010.

[2] G. Tzanetakis and P. Cook, ”Musical genre classification of audio signals,” in IEEE Transactions on Speech and Audio Processing, vol. 10, no. 5, pp. 293-302, July 2002, doi: 10.1109/TSA.2002.800560.

[3] G. Sharma, K. Umapathy and S. Krishnan, ”Trends in audio signal feature extraction methods”, Applied Acoustics, Vol. 158, 2020.

[4] N. Scaringella, G. Zoia and D. Mlynek, ”Automatic genre classification of music content: a survey,” in IEEE Signal Processing Magazine, vol. 23, no. 2, pp. 133-141, March 2006, doi: 10.1109/MSP.2006.1598089.

[5] C. Weihs, U. Ligges, “Mörchen, F. et al. Classification in music research,” ADAC 1, 255–291 (2007). https://doi.org/10.1007/s11634-007-0016-x

[6] Y. Li, X. Zhang, H. Jin, X. Li, Q. Wang, Q. He and Q. Huang. ”Using multi-stream hierarchical deep neural network to extract deep audio feature for acoustic event detection,” Multimed Tools Appl 77, 897–916 (2018).

[7] A. Ghosal, R. Chakraborty, R. Chakraborty, S. Haty, B. C. Dhara and S. K. Saha, ”Speech/Music Classification Using Occurrence Pattern of ZCR and STE,” 2009 Third International Symposium on Intelligent Information Technology Application, Shanghai, 2009, pp. 435-438, doi: 10.1109/IITA.2009.427.

[8] S. Choi, W. Kim, S. Park, S. Yong and J. Nam, ”Korean Singing Voice Synthesis Based on Auto-Regressive Boundary Equilibrium Gan,” ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 7234-7238, doi: 10.1109/ICASSP40776.2020.9053950.

[9] K. Choi, G. Fazekas, M. Sandler and K. Cho, ”Convolutional recurrent neural networks for music classification,” 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, 2017, pp. 2392-2396, doi: 10.1109/ICASSP.2017.7952585.

[10] Y. Li, X. Li, Y. Zhang, W. Wang, M. Liu and X. Feng, ”Acoustic Scene Classification Using Deep Audio Feature and BLSTM Network,” 2018 International Conference on Audio, Language and Image Processing (ICALIP), Shanghai, 2018, pp. 371-374, doi: 10.1109/ICALIP.2018.8455765.

[11] Y. M.G. Costa, L. S. Oliveira, C. N. Silla, “An evaluation of Convolutional Neural Networks for music classification using spectrograms,” in Applied Soft Computing, vol. 52, pp. 28-38, 2017

[12] A. Graves, “Long Short-Term Memory. In: Supervised Sequence Labelling with Recurrent Neural Networks,” Studies in Computational Intelligence, vol. 385. Springer, Berlin, Heidelberg, 2012.

[13] Y. KIKUCHI, N. AOKI and Y. DOBASHI, ”A Study on Automatic Music Genre Classification Based on the Summarization of Music Data,” 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Fukuoka, Japan, 2020, pp. 705-708, doi: 10.1109/ICAIIC48513.2020.9065046.

[14] A. Ghosal, R. Chakraborty, R. Chakraborty, S. Haty, B. C. Dhara and S. K. Saha, ”Speech/Music Classification Using Occurrence Pattern of ZCR and STE,” 2009 Third International Symposium on Intelligent Information Technology Application, Shanghai, 2009, pp. 435-438, doi: 10.1109/IITA.2009.427.

[15] M.I. Mandel, D. Ellis, “Song-Level Features and Support Vector Machines for Music Classification,” ISMIR 2005, 6th International Conference on Music Information Retrieval, London, UK, 11-15 September 2005, Proceedings, 2005

[16] F. A. Gers, J. Schmidhuber and F. Cummins, ”Learning to Forget: Continual Prediction with LSTM,” in Neural Computation, vol. 12, no. 10, pp. 2451-2471, 1 Oct. 2000, doi: 10.1162/089976600300015015.

[17] S. Hochreiter and J. Schmidhuber . ”Long Short-Term Memory.” Neural Computation 9.8(1997):1735-1780.

[18] Kaggle [Online]. Avaliable at: https://www.kaggle.com/andradaolteanu/gtzan-dataset-music-genre-classification

[19] J. H. Foleis, T. F. Tavares, “Texture selection for automatic music genre classification,” Applied Soft Computing Journal 89 (2020)