Mixture of CNN Experts from Multiple Acoustic Feature Domain for Music Genre Classification

Yang Yi, Kuan-Yu Chen and Hung-Yan Gu

National Taiwan University of Science and Technology
E-mail: {m10515801, kychen, guhy}@mail.ntust.edu.tw

***Abstract*—In the field of music information retrieval (MIR), audio spectrogram can carry a great deal of information about the music content so as to be a robust visual representation for music signal. Recently, many research literatures show that convolutional neural network (CNN) has ability to capture indicative acoustic patterns from spectrogram input, and make remarkable performance on MIR-related tasks such as music genre classification (MGC). In this paper, we continue the line of research to explore different types of spectrograms, to emphasize different characteristics of music genre for the MGC task. To jointly leverage all of these features, in this paper, a mixture of experts (MoE) system is proposed. More formally, a set of MGC models can be derived by using the various spectrogram-based statistics. Then we treat each model as an individual expert. Accordingly, a neural mixture model is introduced to collect and compile the predictions from the expert models, and then to output a final decision for a given music to be predicted. In a nutshell, our major contributions in this paper are at least twofold. On one hand, we comprehensively examine several spectrogram-based features for the MGC task. On the other hand, a neural-based MoE system, which can dynamically decide the weighting factor for each expert system, is proposed to enhance the performance of the MGC task [1]. Experimental results demonstrate that the proposed framework not only can achieve success results than individual expert models, but has ability to provide a comparable classification accuracy to the SOTA systems.**

I INTRODUCTION

In the context of MGC task, acoustic features can have a potent effect on the performance. For the past decades, numbers of the research literatures investigate multiple acoustic features which successfully applied to MIR-related tasks. In the conventional study, we extract different kinds of acoustic features from the audio signal by using spectral analysis, to provide robust representation such as octave-based spectral contrast (OSC), audio spectrum envelope (ASE) and Mel-frequency cepstral coefficients (MFCC) [1, 2]. Also, with musical knowledge and machine learning techniques, we can produce higher-level features from music signals like chord [3]. Meanwhile, to deal with the features with high dimensionality such as spectrogram, one of the popular solutions is using PCA to drop the input data into a lower-dimensional space [24, 25], or constructing the codebook utilized by sparse coding algorithms [6]. Therefore, the aforementioned methods provide reduced feature vectors that can efficiently lower the computation burden for classifiers such as SVM [2, 6, 24], GMM [26], and KNN [1]. Because each kind of acoustic feature usually present various aspects of the music signal, such as timbral, rhythmic and pitch, it is reasonable that combining multiple acoustic features can improve MGC accuracy [4]. Besides, with different combination strategies on these acoustic features, the classification accuracy will be various on the MGC task. One of the relevant research literatures has studied both feature-level and decision-level methods for feature combination, the results show that stacked generalization has the best performance [2].

With the rapid growth of deep learning recently, deep neural networks (DNNs) start playing an important role in MIR-related tasks. Convolutional neural network (CNN) is one of the popular architectures, which usually be used in automatic music tagging and the MGC tasks [8-11, 13, 14, 21]. Furthermore, CNN can be a robust feature extractor of spectrogram input. There has been research indicated that each of the CNN layers can find a different level of acoustic patterns from spectrogram input [7], and use activations of feature maps of multiple layers in pre-trained CNN to concatenate as a feature vector, which shows good performance in the MGC task [8]. There is a recent paper shows that even the filters in CNN are randomly weighted, it still can perform reasonably as a feature extractor [9]. Meanwhile, by using such as transfer learning or well-designed network architectures, CNN can gain extra performance improvements in the MGC task [8, 10, 11]. Besides, feature combination is also one of the popular approaches in deep learning structures: There has been used DNN for extracting spectral and temporal features from different source input, an expected performance gain is provided by combination of these two features as input of the classifier [12]; Or train two CNNs by different feature sets, provide better classification accuracy with late fusion strategy than each network themselves [13].

[1] The source code is available at https://github.com/superlyy/apsipa_2019.

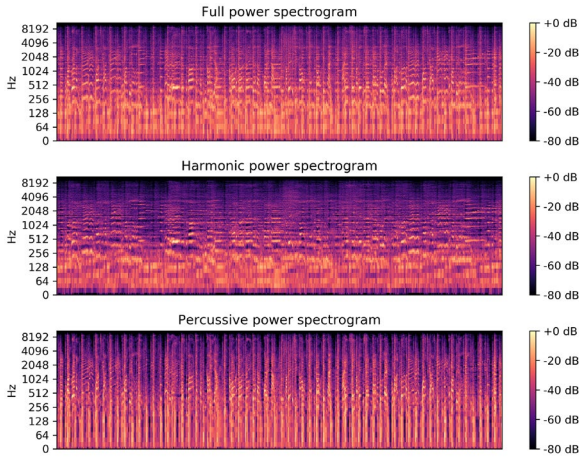

Fig. 1 The log-scaled power spectrogram (upper) with its harmonic (middle) and percussive (lower) spectrogram from a 30-second blues music clip.

Motivated by the well capturing ability of acoustic patterns on CNN, we plan to use different sources of spectrogram as our acoustic features, which include the spectrograms based on harmonic and percussive components, both of the features can provide significant improvement on the MGC task [14]; modulation spectrogram is being used for providing temporal dynamics. Also, we introduce MFCC as baseline feature. In our studied system, each of the acoustic features will train a corresponding CNN, then the well-trained CNN models are treated as experts from multiple acoustic feature domains. We expect that these expert models are designed for extracting different characteristics from the music signal, which means each expert network is able to decide the most probably music genre from the input data through its perspective. After that, three mixer models have been provided in the architecture of mixture of experts. We expect this architecture can provide a reasonable improvement in the MGC task.

II. ACOUSTIC FEATURES

A. Harmonic/Percussive Spectrogram

Harmonic and percussive signal are containing a specific kind of the information about the original music signal, with the support that the shape of the spectral envelope of harmonics and percussion when they have been separated is useful for genre classification [15]. There are serval ways to implement Harmonic Percussive Source Separation (HPSS), one of the approaches by using median filtering shows faster and more effective than the others [16]. Firstly, this approach intuitively accept that stable harmonic or stationary components form horizontal ridges on the spectrogram, while percussive components from vertical ridges with a broadband frequency response. The concept of using median filters individually in the horizontal and vertical directions on the spectrogram which computed from music signal, to separate the spectrograms of harmonic and percussive patterns. Letting $S(f,t)$ represents an original power spectrogram of a given music signal computed by short time Fourier transform (STFT), the result of HPSS can satisfy the condition by

$$S(f,t) = H(f,t) + P(f,t), \tag{1}$$

where $H(\cdot)$ and $P(\cdot)$ represent harmonic power spectrogram and percussive power spectrogram, respectively. f and t denote the index and time frame. Fig. 1 shows a power spectrogram generated from a music clip in log scaled, with its harmonic and percussive power spectrograms.

B. Modulation spectrogram

It is not only spectral features that are important but temporal features also play a key role in improvement for the MGC task. Base on the research from [12], the modulation spectrum is suitable for efficiently analyze the temporal dynamics from the data, and, besides it is simpler to compute than the other approaches. Although the temporal feature used by aforementioned research is based on cepstrogram modulation, we believe that it is sufficient to present investigate the temporal dynamics by modulating on the spectrogram. Given a normalized power spectrogram, which is calculated by its mean value μ_S and standard deviation σ_S as follows:

$$\bar{S}(f,t) = \frac{S(f,t)-\mu_S}{\sigma_S}. \tag{2}$$

Then, the modulation spectrogram is calculated by applying discrete Fourier transform (DFT) on time domain as follows:

$$Mod(f,v) = \left|\sum_{t=0}^{T} \bar{S}(f,t)\, exp\left(-j\frac{2\pi tv}{T}\right)\right|, \tag{3}$$

where v denotes the index of modulation frequency.

C. MFCC

As a popular used acoustic feature, MFCC adopted in many MIR tasks with providing a robust representation. We concatenate 20 MFCC and their first and second-order derivatives in a given time frame, to become our baseline feature.

III. THE PROPOSED METHODOLOGIES

A. Expert Model

There is a MIR-related paper which listed several structures of CNN with a series of comparisons [17], one of the structures called ‘k2c2’, which consists of 5 convolutional layers of 3×3 kernels and max-pooling pooling layer. The experiment results on music tagging task show that ‘k2c2’ has reliable performance especially when the structure parameters are larger than 0.5M. Besides, this structure also has ideal computation time and relatively fewer parameters than the

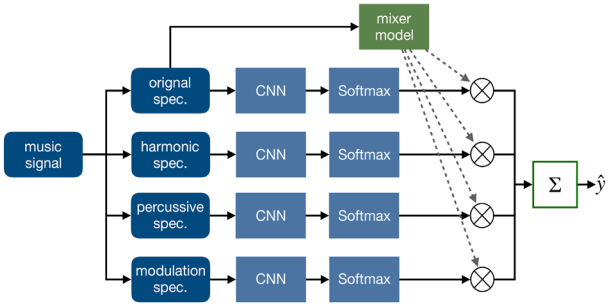

Fig. 2 Overview of the proposed framework.

others. We modify the activation function as Softmax at the output layer of 'k2c2' due to MGC is a single-label classification task, then we use this as our CNN structure. We trained four CNNs with the acoustic features of original, harmonic, percussive and modulation spectrogram, respectively. Consequently, these CNN models are our expert models.

B. Mixture of experts (MoE)

To enhance the performance in the MGC task with expert models, a mixing structure is shown in Fig. 2. Firstly, we calculate four acoustic features from the given music signal for its CNN experts. Then, each of these experts can give a prediction by considering its own feature vector and passing a Softmax activation. The main idea of this strategy is combining expert models' predictions with a desired mixture weight. For a given music signal, we use a mixer model to generate a corresponding weight to each expert model, these weights are satisfied with a constraint as follows:

$$\sum_{m=1}^{M} g_m(x) = 1\,,\ 0 \le g_m(x) \le 1, \tag{4}$$

where x denotes the given music signal of the mixer model, and $g_m(x)$ denotes the weight of m-th expert model. That means this framework can measure how much confidence should put on the specific expert networks based on the input data. And the final decision $\hat{y}$ of the given music signal x can be determined as follows:

$$\hat{y}(x) = \sum_{m=1}^{M} g_m(x) h_m(x), \tag{5}$$

where h_m denotes the prediction of m-th expert model. The studied strategy is basically the same concept as MoE, which has been successfully used in many fields such as language modeling [23]. In our work, we choose the original spectrogram from a given music signal as input feature of the mixer model. We expect that the mixer model can learn the characteristic from the input feature, and find out which of the expert networks will most likely to give the right decision to its music genre. To this end, there are three structures of the mixer model considered:

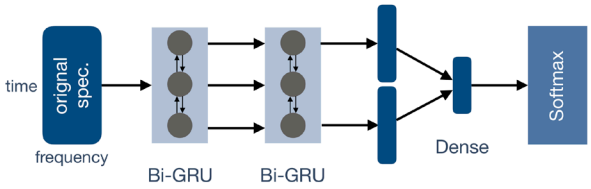

Fig. 3 The proposed RNN-based mixer model.

1. MoE with balanced weight (MoEB): The first structure will be intuitive, which is average the predictions of all the expert networks, or we can say that is combing each expert network with a balanced weight.

2. MoE with CNN as a mixer model (MoEC): We try to use the same structure with the expert model, instead of the output is changed into the number of expert models. We expect that CNN has ability to dynamically decide the weights for the expert CNN models.

3. MoE with RNN as a mixer model (MoER): Recurrent neural network (RNN) with gated recurrent unit (GRU) has been achieved successful results in many sequence modeling problems. In MIR tasks, it can be a good temporal summarizer on acoustic features [18]. We design an RNN structure which is shown in Fig. 3, which contains two layers of bidirectional of GRU (Bi-GRU) units. The second GRU layer outputs the concatenated vector of last hidden states of both directions, then inputs to a dense layer for generating the mixture weights. The time step of RNN input will set to be the number of frames of the spectrograms. We expect that this mixer model can find the temporal pattern from the spectrum of each frame, in order to generate a proper weight for each of the expert models.

For the MoEC and the MoER systems, it should be noted that the expert networks are pre-trained and fixed without fine-tuned with the mixture model.

IV. EXPERIMENTAL SETUP

A. Data Preparation

We perform our experiments on two different datasets:

1. The GTZAN dataset has been very popular in the MIR field [19]. It consists of 1,000 30-second music clips, each clip is annotated with one of 10 genres, the quantity of each genre is balanced. because of the drawbacks such as artist repetition has been indicated [20], we decide to use two different partitioning methods in this dataset. Firstly, we use 10-fold cross-validation with stratified partitioning, which makes each of the folds preserved the percentage of samples for each class. For avoiding the repetition of artists across training, the

TABLE I PERFORMANCE OF THE PROPOSED FRAMEWORK.

	FMA (small)	GTZAN (fault-filtered)	GTZAN (10-fold CV)
original spec.	49.38%	62.07%	81.50%
harmonic spec.	43.38%	61.03%	77.80%
percussive spec.	50.88%	60.00%	79.30%
modulation spec.	55.63%	58.62%	76.70%
MFCC	47.13%	55.52%	78.70%
MoEB	54.13%	65.17%	85.20%
MoEC	55.63%	64.83%	83.80%
MoER	55.88%	66.90%	86.40%

second version called "fault-filtered" of the GTZAN dataset is used [21]. This version of partitioning removes 70 samples, which supported as replicas or distorted waveform, then the remaining samples are manually divided into training (443)/validation (197)/test (290) sets.

2. Like the GTZAN dataset, the FMA dataset provides a sub dataset called FMA small, which contains 8,000 30-second music clips, with 8 balanced genres [22].

B. Feature Configuration

Before computing acoustic features, we resample all the music clips into 22,050Hz, mono signals. When using STFT to compute spectrogram, Hanning window will be chose. we set the window size as 1024 samples, with the hop size of 512 samples, the values are in log-amplitude. We converted spectrogram features with 96 mel frequency bin. Therefore, the input feature matrix size of both original, harmonic and percussive spectrogram will be 96×1292, while due to the symmetry on modulation frequency domain, the size of the modulation spectrogram is set to be 96×646. Each value from the acoustic features is normalized by subtracting mean and divided by standard deviation, where the mean and the standard deviation are calculated from the feature values of each music clip. Both feature extracting and audio processing are implemented with LibROSA [27].

C. Model Configuration

For CNN structure, the number of hidden units of five convolutional layers is (64, 128, 128, 192, 256). Max-pooling is applied after every convolution layer, in order to result the shape of feature maps to 1×1. For RNN structure, the hidden units of Bi-GRU layers and Dense layer are (64, 64, 32). Besides, batch normalization and ELU activation function are used in all the convolutional layers and dense layers. Both CNN and RNN are using Adam as the optimizer and cross-entropy as the loss function. Each of the networks is trained over 50 epochs with the batch size of 16 in every epoch. All the models are built with Keras.

TABLE II COMPARISON TO PREVIOUS STATE-OF-THE-ARTS.

	FMA (small)	GTZAN (fault-filtered)	GTZAN (10-fold CV)
2D CNN [21]	-	63.20%	-
temporal feature [12]	-	65.90%	85.00%
transfer learning [8]	-	-	89.80%
multi-level and multi-scale [10]	-	72.00%	-
artist label [11]	56.87%	72.03%	-
The Proposed MoER	55.88%	66.90%	86.40%

V. RESULTS AND DISCUSSIONS

The experimental results of each expert model (includes MFCC as baseline) and different MoE structures in each dataset or partitioning methods are shown in Table I. As we can see, both of the expert models deliver vary but reasonable performance. Overall results show that GTZAN (10-fold CV) can obtain the highest accuracy. Compare to another different partitioning method, the accuracy of fault-filtered version decreased in general, one of the main reasons is that the 10-fold CV version has the repetition of artist across its training and testing sets, which can make the models be able to capture similarity from music clips by their generated features. In the performance of each expert model, firstly, the expert model based on original the spectrogram serves a high and stable classification accuracy across all the datasets, one of the reasons is that most of the information on the music signal is preserved in the original spectrogram. Then the following part will be harmonic and percussive ones. We believe that these two separated components from the original spectrogram can be experts in conveying certain characteristics of music signal, which help decide its music genre. Although in FMA small, harmonic's expert show relatively low result of 43.38%, and percussive one is even higher than original spectrogram, part of the reason is that the music genre such as electronic, hip-hop, pop or rock, which can be shown more discriminating in their percussive patterns, while the harmonic patterns will be confused with those genres. Besides, the expert model based on modulation spectrogram achieves the highest accuracy of 55.63% in FMA small among the other expert models, this result is even better than two of the MoE structures, a possible reason is that many of the music genres in FMA small can be well explained by their temporal dynamics. Both of the aforementioned expert models can make a comparable performance to the MFCC one. Although MFCC is known to be a robust feature and relatively low data size, but with the help of CNN, many of the acoustic patterns can be extracted from those spectrogram inputs, which are helpful for containing a better result. From the results of proposed MoE structures, we can see that most of them make a noticeable improvement from the proposed expert models. MoEB implements a voting-like approach and serves as a baseline in

MoE, the result proves that the ensemble of expert models has the ability to make a more stable prediction. Compare to MoEB, MoEC shows a performance degradation on GTZAN, we give the possible explanations are the insufficient data size for training the mixer model, or CNN will be hard to learn a proper way to generate the mixture weights based on the feature input. However, MoER can deliver the best performance in all of our experiments, the result indicated that the mixture weights are meaningful, and the temporal information learned by RNN which is helpful for mixer model to generate accurate weights.

Finally, we make a comparison between the proposed framework with some state-of-the-arts in the MGC task. From Table II, our proposed approach can show better results than [12, 21]. Compare to [8, 10, 11], however, we still have some performance gaps between them. Part of the reason is that these researches use transfer learning techniques which pre-train their model on a large dataset such as Million Song Dataset to be their source task, which can effectively enhance the generalization ability of the models.

VI. CONCLUSIONS

In this paper, we explore various spectrogram-based acoustic features, explore different types of spectrograms, such as the harmonic and percussive components generated from the original spectrogram, to emphasize different characteristics of music genre for the MGC task. Besides, modulation from time domain of original spectrogram is also used, which is containing temporal dynamics of music signal. Meanwhile, an MoE system based on acoustic feature domain is studied on the MGC task. The experiments show that the system can make comparable improvement in the MGC task. In the future, we plan to investigate how effective that acoustic feature can specify certain music genres, and study more kinds of acoustic features.

ACKNOWLEDGEMENTS

This work is supported by the Ministry of Science and Technology (MOST) in Taiwan under grant MOST 108-2636-E-011-005 (Young Scholar Fellowship Program), and by the Project J367B83100 (ITRI) under the sponsorship of the Ministry of Economic Affairs, Taiwan.

REFERENCES

[1] C.H. Lee, J.L. Shih, K.M. Yu and H.S. Lin, "Automatic music genre classification based on modulation spectral analysis of spectral and cepstral features," *IEEE Transactions on Multimedia,* pp. 11.4: 670-682, 2009.

[2] Z. Fu, G. Lu, K.M. Ting, D. Zhang, "On feature combination for music classification," in *Proc. of Joint IAPR International Workshops on SPR and SSPR*, pp. 453-462, 2010.

[3] H.T. Cheng, Y.H. Yang, Y.C. Li, I.B. Liao and H.H. Chen, "Automatic chord recognition for music classification and retrieval," in *Proc. of IEEE International Conference on Multimedia and Expo*, pp. 1505-1508, 2008.

[4] B.K. Baniya, D. Ghimire, and J. Lee, "Evaluation of different audio features for musical genre classification," in *Proc. of SiPS on IEEE*, pp. 260-265, 2013.

[5] J. Shen, J. Shepherd and A. H.H. Ngu, "Towards effective content-based music retrieval with multiple acoustic feature combination," *IEEE Transactions on Multimedia*, pp. 8.6: 1179-1189, 2006.

[6] C. C. M. Yeh, L. Su, and Y. H. Yang, "Dual-layer bag-of-frames model for music genre classification," *in Proc. of ICASSP*, pp. 246-250, 2013.

[7] K. Choi, G. Fazekas, and M. Sandler, "Explaining deep convolutional neural networks on music classification," arXiv:1607.02444, 2016.

[8] K. Choi, G. Fazekas, M. Sandler and K. Cho, "Transfer learning for music classification and regression tasks," in *Proc. of ISMIR*, 2017.

[9] J. Pons, and X. Serra, "Randomly weighted CNNs for (music) audio classification," in *Proc. of ICASSP*, 2019, pp. 336-340.

[10] J. Lee and J. Nam, "Multi-level and multi-scale feature aggregation using pretrained convolutional neural networks for music auto-tagging," in *IEEE Signal Processing Letters*, pp. 1208-1212, 2017.

[11] J. Park, J. Lee, J. Park, J. W Ha and J. Nam, "Representation learning of music using artist labels," in *Proc. of ISMIR*, 2018.

[12] I.Y. Jeong and K. Lee, "Learning temporal features using a deep neural network and its application to music genre classification," in *Proc. of ISMIR*, pp. 434-440, 2016.

[13] C. Senac, T. Pellegrini, F. Mouret and J. Pinquier, "Music feature maps with convolutional neural networks for music genre classification," in *Proc. of International Workshop on Content-Based Multimedia Indexing, ACM*, pp. 19, 2017.

[14] G. Gwardys and D. Grzywczak, "Deep image features in music information retrieval," *International Journal of Electronics and Telecommunications*, pp. 60.4: 321-326, 2014.

[15] H. Rump, S. Miyabe, E. Tsunoo, N. Ono, S. Sagayama, "Autoregressive MFCC models for genre classification improved by harmonic-percussion separation," in *Proc. of ISMIR*, 2010.

[16] D. Fitzgerald, "Harmonic/percussive separation using median filtering," in *Proc. of DAFX10*, 2010.

[17] K. Choi, G. Fazekas, M. Sandler, K. Cho, "Convolutional recurrent neural networks for music classification," in *Proc. of ICASSP*, 2017.

[18] K. Choi, G. Fazekas, K. Cho and M. Sandler, "A tutorial on deep learning for music information retrieval," arXiv:1709.04396, 2017.

[19] G. Tzanetakis and P. Cook, “Musical genre classification of audio signals,” *IEEE Transactions on speech and audio processing*, pp. 10.5: 293-302, 2002.

[20] B.L. Sturm, “The GTZAN dataset: its contents, its faults, their effects on evaluation, and its future use,” arXiv:1306.1461, 2013.

[21] C. Kereliuk, B.L. Sturm and J. Larsen, “Deep learning and music adversaries,” *IEEE Transactions on Multimedia*, 2015.

[22] M Defferrard, K Benzi, P Vandergheynst and X. Bresson, “Fma: a dataset for music analysis,” arXiv:1612.01840, 2016.

[23] K. Irie, S. Kumar, M. Nirschl and H. Liao, “RADMM: recurrent adaptive mixture model with applications to domain robust language modeling,” in *Proc. of ICASSP*, 2018.

[24] Y. Panagakis, C. Kotropoulos and G.R. Arce, “Non-negative multilinear principal component analysis of auditory temporal modulations for music genre classification,” *IEEE Transactions on Audio, Speech, and Language Processing*, 18.3: 576-588, 2009.

[25] A. van den Oord, S. Dieleman and B. Schrauwen, “Transfer learning by supervised pre-training for audio-based music classification,” in *Proc. of ISMIR*, 2014.

[26] F.A. de Leon, K. Martinez, "Music genre classification using polyphonic timbre models," in *Proc. of International Conference on Digital Signal Processing*, 2014

[27] B. McFee, C. Raffel, D. Liang, D.P. Ellis, M. McVicar, E. Battenberg, et al., “librosa: Audio and music signal analysis in python,” in *Proc. of the 14th python in science conference*, 2015.