HIGH-LEVEL MUSIC DESCRIPTOR EXTRACTION ALGORITHM BASED ON COMBINATION OF MULTI-CHANNEL CNNS AND LSTM

Ning Chen
East China University of Science and Technology
nchen@ecust.edu.cn

Shijun Wang
East China University of Science and Technology
sqwsj@hotmail.com

ABSTRACT

Although Convolutional Neural Networks (CNNs) and Long Short Term Memory (LSTM) have yielded impressive performances in a variety of Music Information Retrieval (MIR) tasks, the complementarity among the CNNs of different architectures and that between CNNs and LSTM are seldom considered. In this paper, multi-channel CNNs with different architectures and LSTM are combined into one unified architecture (Multi-Channel Convolutional LSTM, MCCLSTM) to extract high-level music descriptors. First, three channels of CNNs with different shapes of filter are applied on each spectrogram image chunk to extract the pitch-, tempo-, and bass-relevant descriptors, respectively. Then, the outputs of each CNNs channel are concatenated and then passed through a fully connected layer to obtain the fused descriptor. Finally, LSTM is applied on the fused descriptor sequence of the whole track to extract its long-term structure property to obtain the high-level descriptor. To prove the efficiency of the MCCLSTM model, the obtained high-level music descriptor is applied to the music genre classification and emotion prediction task. Experimental results demonstrate that, when compared with the hand-crafted schemes or conventional deep learning (Multi Layer Perceptrons (MLP), CNNs, and LSTM) based ones, MCCLSTM achieves higher prediction accuracy on three music collections with different kinds of semantic tags.

1. INTRODUCTION

The amount of online music tracks is constantly growing, which makes it difficult to tag them manually. Without accurate labels, most of the tracks cannot be accessed. So, auto-tagging technique has become a hot topic in the field of Music Information Retrieval (MIR) for the past two decades. It can be used in music classification, music retrieval, and music recommendation systems.

In the past ten years, deep learning models, such as Deep Neural Networks (DNNs), Convolutional Neural Networks (CNNs), and Long Short Term Memory (LSTM) [9] have achieved tremendous success for a variety of MIR tasks, such as onset detection [20], emotion recognition [13], chord estimation [5], rhythm stimuli recognition [22], auto-tagging [3, 16], source separation [10], or music recommendation [25], etc. It has been proved that deep learning based models are superior to hand-crafted ones in music content analysis because [16]: i) The nonlinear mapping in deep learning model (e.g. CNNs) is suitable for describing the time-varying nonlinear property of music signal. ii) The hierarchical architecture of deep learning model is fit for representing the hierarchical nature of music in both time domain (onset, rhythm) and frequency domain (note, chord) [16]. iii) Long-term dependencies property of music (music structure or recurrent harmonies), which is important for human music perception and understanding, can be modeled by deep learning model (e.g. LSTM) very well [11].

Despite the rich potential of CNNs and LSTM in describing music properties, they are individually limited in their modeling capability [19]. In [16], the CNNs were adopted to learn high-level descriptor from the spectrogram image of the music signal, and the filter shape of CNNs was studied to make it suitable for representing different music relevant descriptors. It was verified that wider filters and higher filters may be capable of learning longer temporal dependencies and more spread timbral features, respectively. This scheme achieved competitive results in auto-tagging on the Ballroom dataset [8]. However, as shown in [16, 19], CNNs may only model the local context, such as instrument's timbre or musical units, well, but not the long-term dependencies, such as music structure or recurrent harmonies, of the music. As for the LSTMs based schemes, their main issues are two aspects. On the one hand, the temporal modeling is usually done on the low-level descriptor, which makes it difficult to disentangle underlying factors of variation within the input [12]. On the other hand, as shown in [15], there is no intermediate nonlinear hidden layer in LSTM, so the history of previous inputs cannot be summarized efficiently.

To take advantage of the complementarity between CNNs and LSTM, some researchers proposed to combine them in a unified architecture [2, 4]. In [2], LSTM

and CNNs are combined in parallel to exploit sequential correlation and local spectro-temporal information. Then, the outputs of the CNNs and LSTM are combined by the fully connected layers to obtain the fused descriptor. Experimental results demonstrated that this scheme outperformed the conventional DNNs, CNNs, or LSTM based ones in acoustic scene classification task. In [19], considering that: "CNNs are good at reducing frequency variations, LSTMs are good at temporal modeling, and DNNs are appropriate for mapping features to a more separable space", the three models are combined into one unified architecture to take advantage of the complementarity among them. This scheme achieved better performances than the LSTM based one in the voice search task.

In this paper, a new deep learning based architecture (called Multi-Channel Convolutional LSTM, MCCLSTM) is proposed for high-level music descriptor extraction. MCCLSTM is different from the methods discussed above as it takes advantage of the complementarity among CNNs with different architectures and that between CNNs and LSTM in modeling music properties. Considering that different musical properties correspond to different time-frequency resolutions [16], the descriptor extracted by one CNN with a specific filter may not characterize the music property comprehensively. So, in the proposed scheme, three channels of CNNs with different shapes of filters are adopted to extract pitch-, tempo-, and bass-relevant descriptors from the spectrogram image, respectively. Then, the outputs of each CNNs are concatenated and then passed trough a fully connected layer to map to the fused descriptor. Finally, since it has been verified in [15] that the performance of LSTM can be improved greatly when provided with high-level descriptor, the LSTM is put as a higher level of the fully connected layer in the proposed scheme to learn the time dependency in the fused descriptor sequence to extract the long-term structure of the whole track. Since the obtained high-level descriptor contains both local context based and long-term structure based information of the music, it may describe the music property more comprehensively. Experimental results demonstrate that the proposed model is superior to the hand-crafted schemes [14, 18] and the conventional deep learning (Multi Layer Perceptrons (MLP) [21], CNNs [16], and LSTM) based ones in music auto-tagging task on three music collections with different kinds of semantic tags.

The rest of this paper is organized as follows. The proposed scheme is described in detail in Section 2. The performances of the proposed scheme in music auto-tagging task in comparison with other state-of-the-art schemes are evaluated and discussed in Section 3. Conclusions and prospects on future work are given in Section 4.

2. MCCLSTM MODEL

The MCCLSTM architecture is shown in Figure 1.

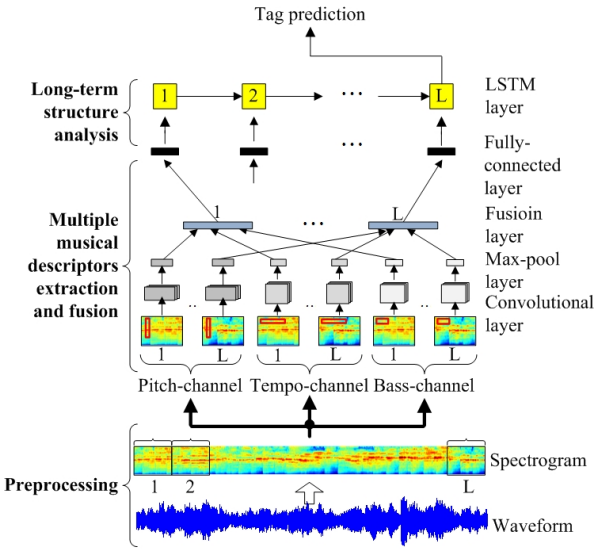

Figure 1: Multi-Channel Convolutional LSTM (MCCLSTM) architecture.

2.1 Preprocessing

The same preprocessing procedure shown in [16] is adopted in the proposed scheme. First, the Short-Time Fourier Transform (STFT) is applied to the input music audio signal, whose sampling rate is 44100 Hz, to obtain the spectrum of it. In STFT, a Blackman-Harris window of 2048 samples is chosen, and the hop size is 1024 samples. Next, the 40-band Mel filter-bank is applied on the obtained spectrum to generate the corresponding spectrogram image of it. Then, the whole spectrogram image is split into L chunks without overlapping. The size of each spectrogram chunk is $M \times N$, where M and N stand for the number of frequency bins and that of frames, respectively.

2.2 Multiple Musical Descriptors Extraction and Fusion

There is no one universal deep learning architecture or hand-crafted scheme that performs well in modeling multiple music properties at the same time. To solve this problem and describe the music content more comprehensively, three channels of CNNs with different shape of filters are adopted and combined in the proposed scheme to obtain fused descriptor, which contains tempo-, pitch-, and bass-relevant information of the input music. This idea was first proposed in [16] to combine the tempo- and pitch-relevant information and was modified in this paper by adding another bass-relevant information.

- Pitch-channel CNNs: in this channel, a $m \times 1 (m \ll M)$ frequency filter is chosen. This type of filter is designed for modeling frequency features. The upper layer can represent some temporal dependencies from the resulting activations as well [16]. In the proposed scheme, this channel of CNNs are responsible for extracting pitch, timbre, or

equalization setups relevant descriptor of the music.

- Tempo-channel CNNs: in this channel, a $1 \times n (n \ll N)$ temporal filter is adopted. This kind of filter will be suitable for learning temporal dependencies but not frequency dependencies. Also, the upper layer may exploit the frequency relations [16]. In the proposed scheme, this channel of CNNs try to learn rhythmic/tempo relevant patterns of the music.

- Bass-channel CNNs: in this channel, a $m \times n (m \ll M, n \ll N)$ filter is taken. This type of filter is capable of learning time and frequency features at the same time. Different musical aspects can be learned by such filters with different combination of m and n. Considering that the task of this channel of CNNs is to model the bass- or kick-relevant feature, which most entails finding changes over time, a filter that is wide in time and narrow in frequency (i.e. $m < n$) is adopted [16].

To take advantage of the complementarity among the obtained pitch-, tempo-, and bass-relevant descriptors, they are concatenated on the fusion layer and then passed through a fully connected layer to generate the fused descriptor.

2.3 Long-Term Structure Analysis

Although CNNs may model the local context in the spectrogram chunk well, it may not model the long-term structure (music structure or recurrent harmonies) of the whole track, which is quite important for human music perception and understanding [11]. To solve this problem, a LSTM layer is added on top of the fully connected layer. It will help to learn the time dependence in the fused descriptor sequence of the whole track. The number of the nodes in LSTM is equal to that of the chunks included in the whole track. Since the obtained high-level descriptor contains the information of different music properties (pitch, tempo, and bass) and that of long-term structure of the music as well, it may be more suitable for auto tagging of music with different kinds of semantic tags.

3. EXPERIMENTS

In the experiment, the input of the architecture is 40-dimensional log-mel filterbank based spectrogram images, computed every 3.712s (i.e., $M = 40, N = 80$). As shown in Figure 1, all three CNNs channels are composed of 1 convolutional layer and 1 max-pooling layer. The size of each kernel in the multi-channel convolutional neural network is listed in Table 1. When 2 (or 3) CNNs channels are fused, the fully connected layer and the LSTM layer contain 200 (or 400) and 100 (or 200) units, respectively. The number of nodes in LSTM is equal to that of the chunks included in the whole spectrogram. The weights for all CNN and LSTM layers are randomly initialized to be Gaussian, with a variance of 1. And a softmax layer is added upon the LSTM layer to discriminate the tag of the overall input music. Six kinds of architectures based on different combinations of CNNs and LSTM (see Table 2) are studied in the experiment. To verify the efficiency of the proposed high-level musical descriptor extraction scheme, its performances in music auto-tagging task are tested on three music collections with different kinds of semantic tags, in comparison with those obtained by the hand-crafted schemes or conventional deep learning-based ones.

The whole architecture shown in Figure 1 is trained together with the categorical cross-entropy criterion, using the asynchronous stochastic gradient descent optimization strategy. The weights for all CNN and LSTM layers are randomly initialized to be Gaussian, with a variance of 1. The prediction accuracies obtained by deep learning-based schemes are computed using 10-fold cross validation with a randomly generated train-validation-test split of 80%-10%-10%.

Layer name	P-channel	T-channel	B-channel
Convolutional layer	(32,1)	(1,60)	(13,9)
Max-pooling layer	(1,80)	(40,1)	(4,4)

Table 1: Size of each kernel in the multi-channel convolutional neural network.

ID	Architecture
R0	CNNs (T)
R1	LSTM only
R2 [16]	CNNs (T+P)
R3	CNNs (T+P+B)
R4	CNNs (T+P)+LSTM
MCCLSTM	CNNs (T+P+B)+LSTM

Table 2: Six architectures studied in the experiments.

3.1 Datasets

The following three music collections with different kinds of semantic tags are adopted to test the performances of the proposed scheme in auto-tagging task.

- GTZAN genre collection [24]: Although GTZAN dataset suffers from some repetitions, mislabelings and distortions problems [17], it is often adopted to evaluate genre classification accuracy. This dataset is composed of 1000 audio tracks, each 30 seconds long. It contains 10 genres (blues, classical, country, disco, hiphop, jazz, metal, pop, reggae, and rock), each of which is represented by 100 tracks.

- Ballroom dataset: This dataset comprises 698 audio tracks, each around 30 seconds long. It contains 8 ballroom dancing genres (cha-cha-cha 111, jive 60, quickstep 82, rumba 98, samba 86, tango 86, viennese valtz 65, and slow waltz 110).

- Soundtracks dataset [7] for music and emotion: This dataset contains 470 film music excerpts, each 15-30 seconds long. The tag of each excerpts is one of the five discrete emotions: anger 61, fear 116, sadness 108, happiness 89, and tenderness 96 [6].

3.2 Experimental Results

To verify the efficiency of the MCCLSTM model in music auto-tagging task, its performance, in terms of prediction accuracy, is compared with those obtained by the conventional hand-crafted schemes and other deep learning-based ones (MLP, CNNs, and LSTM) on each of above datasets.

3.2.1 Baselines

In the experiment, as shown in Table 3-5, two hand-crafted schemes [14, 18] and four deep learning-based ones (MLP [21], R0, R1, and CNNs-3) are included as baselines. In R0, the output of the tempo-channel CNNs is used for tag prediction, directly. While, in R1, the LSTM is learned directly on the spectrogram chunk sequence. The CNNs-3 scheme is composed of two convolutional levels, two max-pooling levels, and one fully connected level (200 units). The sizes of the two convolutional layers are (5,5) and (3,3), respectively. The sizes of the two max-pooling layers are both (2,2). The two max-pooling layers are alternated with convolutional layers. It should be noted that, since the codes of the schemes in [14], [18], and [21] are not available, we just include the prediction accuracies obtained by them on specific music collections. As shown in Table 3-5, when compared with the hand-crafted schemes [14, 18], the MCCLSTM scheme can enhance the prediction accuracy from 3.50% to 16.95%. As for the deep learning-based ones (MLP [21], R0, R1, and CNNs-3), MCCLSTM scheme can achieve a higher prediction accuracy of 1.69%-32.99%, 4.90%-27.90%, and 9.85%-22.35%, on GTZAN, Ballroom, and Soundtracks datasets, respectively. So, it is verified that the proposed scheme is superior to the hand-crafted schemes and conventional deep learning-based ones in auto tagging task across the datasets included.

Schemes	Accuracy: mean%±std
[14]	72.80
[21] (3 layers)	83.00 ± 1.10
CNNs-3	79.80 ± 1.70
R0	51.70 ± 2.60
R1	59.30 ± 2.20
R2 [16]	78.40 ± 1.90
R3	82.90 ± 2.11
R4	83.70 ± 1.10
MCCLSTM	84.69 ± 1.76

Table 3: Performance comparison on GTZAN dataset.

Schemes	Accuracy: mean%±std
[14]	88.40
CNNs-3	87.00 ± 1.32
R0	81.79 ± 4.72
R1	64.00 ± 1.50
R2 [16]	87.68 ± 4.44
R3	89.45 ± 2.18
R4	90.32 ± 1.12
MCCLSTM	91.90 ± 2.33

Table 4: Performance comparison on Ballroom dataset.

Schemes	Accuracy: mean%±std
[18]	57.40 ± 5.50
CNNs-3	60.87 ± 3.76
R0	52.00 ± 2.20
R1	64.50 ± 2.00
R2 [16]	57.28 ± 2.85
R3	63.04 ± 2.16
R4	73.70 ± 2.47
MCCLSTM	74.35 ± 1.63

Table 5: Performance comparison on Soundtracks dataset.

3.2.2 Multi-Channel CNNs Based Schemes

In [16] (denoted as R2 in this paper), the outputs of the pitch-channel CNNs and the tempo-channel CNNs in Figure 1 were concatenated to obtain the fused musical descriptor, which then contains both pitch- and tempo-relevant information. To make the fused descriptor contain bass relevant information also, a bass-channel CNNs is added in R2 to obtain the three-channel CNNs based one (denoted as R3). As shown in Table 3-5, R3 achieves higher prediction accuracy than [16] on all three datasets. Especially, for the music mood auto-tagging (see Table 5), R3 enhances the prediction accuracy by 5.76% when compared with [16]. The latent reason may be that the bass relevant information plays a crucial role in mood classification [1].

3.2.3 Multi-Channel CNNs + LSTM Based Schemes

Music can be described as sequences of events that are structured in pitch and time [23]. So, how to learn and represent such complex event sequences (or long-term structure) is very important for music perception and cognition. However, CNNs may only model the local context well but not the long-term dependencies contained in the whole track [16], which will affect the accurate describing of the music properties. Considering that LSTM is good at extracting the sequential information from the consecutive features, a LSTM layer is added on the top of the fully connected layer (as shown in Figure 1) to model the long-term structure property of the music. To show the benefits of LSTM, it is applied on the two-channel and three-channel CNNs fused descriptor, respectively, to construct R4 and MCCLSTM schemes. The experimental results shown in Table 3-5 indicate that for the two-channel CNNs based scheme (R2), the introducing of LSTM layer can help to enhance the prediction accuracy of 5.30%, 2.64%, and 16.42% on GTZAN, Ballroom, and Soundtracks, respectively. For the three-channel CNNs based scheme (R3), the adding of LSTM layer can help to enhance the prediction accuracy of 1.79%, 2.45%, and 11.31% on GTZAN, Ballroom, and Soundtracks, respectively. So, it is verified that adopting LSTM to analyze the time dependencies contained in the fused descriptor sequence further may help to describe the musical characteristic more accurately and comprehensively.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we present a unified deep learning architecture (MCCLSTM) for high-level musical descriptor extraction. First, the CNNs with different resolutions are utilized to analyze each spectrogram chunk of the input music to extract different music property-relevant descriptors, respectively. Then, the outputs of each CNNs channels are concatenated and then passed through a fully connected layer to obtain the fused descriptor. Finally, the LSTM is performed on the fused descriptor sequence to model the long-term structure of the input music. To verify the efficiency of the MCCLSTM scheme, its performance in music auto tagging are compared with those obtained by the hand-crafted schemes and the conventional deep learning (MLP, CNNs, and LSTM) based ones on three music collections with different kinds of semantic tags. Experimental results demonstrate that the proposed scheme is superior to hand-crafted schemes in [14, 18] and other deep learning-based ones ([16, 21], R0, R1, R3, R4, and CNNs-3). However, since the fused descriptor is obtained by concatenating the outputs of each CNNs channel, the complementarity among these three descriptors cannot be fused efficiently. So, our future work is to study new fusion mechanism, which can utilize the common as well as complementary aspects of each musical descriptors more efficiently.

5. ACKNOWLEDGEMENT

This work is supported by the National Natural Science Foundation of China (No. 61271349).

6. REFERENCES

[1] Jakob Abeßer, Hanna Lukashevich, Christian Dittmar, and Gerald Schuller. Genre classification using bass-related high-level features and playing styles. In *Proceedings of 10th International Conference on Music Information Retrieval (ISMIR 2009)*, pages 453–458, 2009.

[2] Soo Hyun Bae, Inkyu Choi, and Nam Soo Kim. Acoustic scene classification using parallel combination of lstm and cnn. In *2016 Detection and Classification of Acoustic Scenes and Events (DCASE 2016)*, pages 1–5, 2016.

[3] Keunwoo Choi, George Fazekas, and Mark Sandler. Automatic tagging using deep convolutional neural networks. Proceedings of 17th International Conference on Music Information Retrieval (ISMIR 2016), 2016.

[4] Keunwoo Choi, György Fazekas, Mark Sandler, and Kyunghyun Cho. Convolutional recurrent neural networks for music classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2392–2396. IEEE, 2017.

[5] Junqi Deng and Yu-Kwong Kwok. A hybrid gaussian-hmm-deep-learning approach for automatic chord estimation with very large vocabulary. Proceedings of 17th International Conference on Music Information Retrieval (ISMIR 2016), 2016.

[6] Tuomas Eerola and Jonna K Vuoskoski. A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music*, 2010.

[7] Tuomas Eerola and Jonna K Vuoskoski. A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music*, 39(1):18–49, 2011.

[8] Fabien Gouyon, Anssi Klapuri, Simon Dixon, Miguel Alonso, George Tzanetakis, Christian Uhle, and Pedro Cano. An experimental comparison of audio tempo induction algorithms. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1832–1844, 2006.

[9] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[10] Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson, and Paris Smaragdis. Joint optimization of masks and deep recurrent neural networks for monaural source separation. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 23(12):2136–2147, 2015.

[11] Eric J Humphrey, Juan P Bello, and Yann LeCun. Feature learning and deep architectures: new directions for music informatics. *Journal of Intelligent Information Systems*, 41(3):461–481, 2013.

[12] Bernhard Lehner, Gerhard Widmer, and Sebastian Bock. A low-latency, real-time-capable singing voice detection method with lstm recurrent neural networks. In *2015 23rd European Signal Processing Conference (EUSIPCO)*, pages 21–25. IEEE, 2015.

[13] Xinxing Li, Haishu Xianyu, Jiashen Tian, Wenxiao Chen, Fanhang Meng, Mingxing Xu, and Lianhong Cai. A deep bidirectional long short-term memory based multi-scale approach for music dynamic emotion prediction. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 544–548. IEEE, 2016.

[14] Athanasios Lykartsis and Alexander Lerch. Beat histogram features for rhythm-based musical genre classification using multiple novelty functions. In *Proceedings of 16th International Conference on Music Information Retrieval (ISMIR 2015)*, pages 434–440, 2015.

[15] Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. How to construct deep recurrent neural networks. In *Proceedings of the Second International Conference on Learning Representations (ICLR 2014)*, 2014.

[16] Jordi Pons, Thomas Lidy, and Xavier Serra. Experimenting with musically motivated convolutional neural networks. In *2016 14th International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 1–6. IEEE, 2016.

[17] Francisco Rodrıguez-Algarra, Bob L Sturm, and Hugo Maruri-Aguilar. Analysing scattering-based music content analysis systems: Wheres the music? In *Proceedings of 17th International Conference on Music Information Retrieval (ISMIR 2016)*, 2016.

[18] Pasi Saari, Tuomas Eerola, and Olivier Lartillot. Generalizability and simplicity as criteria in feature selection: Application to mood classification in music. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(6):1802–1812, 2011.

[19] Tara N Sainath, Oriol Vinyals, Andrew Senior, and Haşim Sak. Convolutional, long short-term memory, fully connected deep neural networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4580–4584. IEEE, 2015.

[20] Jan Schlüter and Sebastian Böck. Improved musical onset detection with convolutional neural networks. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6979–6983. IEEE, 2014.

[21] Siddharth Sigtia and Simon Dixon. Improved music feature learning with deep neural networks. In *Proceedings of 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6959–6963. IEEE, 2014.

[22] Sebastian Stober, Daniel J Cameron, and Jessica A Grahn. Using convolutional neural networks to recognize rhythm stimuli from electroencephalography recordings. In *Advances in neural information processing systems*, pages 1449–1457, 2014.

[23] Barbara Tillmann. Music and language perception: expectations, structural integration, and cognitive sequencing. *Topics in cognitive science*, 4(4):568–584, 2012.

[24] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, 10(5):293–302, 2002.

[25] Aaron Van den Oord, Sander Dieleman, and Benjamin Schrauwen. Deep content-based music recommendation. In *Advances in neural information processing systems*, pages 2643–2651, 2013.