Long Short-term Memory Recurrent Neural Network based Segment Features for Music Genre Classification

Jia Dai[1], *Shan Liang*[1], *Wei Xue*[1], *Chongjia Ni*[2], *Wenju Liu*[1]

[1]NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[1]University of Chinese Academy of Sciences, Beijing, China
[2]Shandong University of Finance and Economics, shandong, China
[1]{jia.dai,sliang}@nlpr.ia.ac.cn, xuewei.x@gmail.com, cjni_sd@sdufe.edu.cn,lwj@nlpr.ia.ac.cn

Abstract

In the conventional frame feature based music genre classification methods, the audio data is represented by independent frames and the sequential nature of audio is totally ignored. If the sequential knowledge is well modeled and combined, the classification performance can be significantly improved. The long short-term memory(LSTM) recurrent neural network (RNN) which uses a set of special memory cells to model for long-range feature sequence, has been successfully used for many sequence labeling and sequence prediction tasks. In this paper, we propose the LSTM RNN based segment features for music genre classification. The LSTM RNN is used to learn the representation of LSTM frame feature. The segment features are the statistics of frame features in each segment. Furthermore, the LSTM segment feature is combined with the segment representation of initial frame feature to obtain the fusional segment feature. The evaluation on ISMIR database show that the LSTM segment feature performs better than the frame feature. Overall, the fusional segment feature achieves 89.71% classification accuracy, about 4.19% improvement over the baseline model using deep neural network (DNN). This significant improvement show the effectiveness of the proposed segment feature.

Index Terms: Long short-term memory, recurrent neural network, music genre classification, scattering transform

1. Introduction

Music is an important kind of audio data, and the study of automatic music classification is an important branch of audio classification. The music genre is one of the most popular ways to describe the music content, and it has wide applications in music information retrieval, music recommendation and online music access.

Some methods have been studied to classify the music genres [1, 2]. Although these methods can achieve the satisfactory performance on small size and moderate size datasets, they can not perform well when the amount of music data is increasing large. The develop of deep neural network (DNN) enables the training for big data [3–6]. Due to the large amount of music tracks on the Internet, the music genre classification using DNN has become a trend [7, 8]. However, there still exists a problem. DNN or other models train the model with independent input frames and can not capture the temporal dependencies of audio sequence. For example, the music tracks have the melody, which can be seen as a special kind of sequence representation, and it is important in distinguishing different categories of music. However, these frame training based models will ignore the sequence information during training. Although there are some studies on the sequence representation of speech data [9, 10], as the sequential characteristics of the speech and music are inherently different, they can not be efficiently applied to music data.

The recurrent neural network (RNN) is a special kind of neural network, which are used for a variety of sequence-labeling tasks [11]. But standard RNN can only make use of the previous limited context. It has limited storage to deal with long sequences because of the problem of vanishing and exploding gradients, hence they have difficultly in learning the long-term dependencies [12]. A solution to this problem is offered by the long short-term memory (LSTM) RNN, which avoids the problematic non-linearity in the recursion [13]. LSTM RNN have achieved the state-of-the-art performance on the sequence discriminative training. The LSTM uses memory blocks to model temporal dependencies, which allows it to more effectively exploit the long-range context than RNN [14]. Each memory block contains self-connected memory cells and three gate units: the input, forget, and output gates. The memory cell uses the three gates to attenuate the input, recurrent, and output signals respectively [13], and can respectively provide write, read, reset operations to the cell. These allow the model to ignore the unimportant inputs, memories, and outputs [12].

In this paper, we propose the LSTM RNN based segment features for music genre classification. In order to get the representation which contains sequence information, we first learn the frame representation of music data using LSTM RNN. The LSTM RNN is trained on the independent frame feature, and then the soft-max probability is donated as the LSTM frame representation of music data. As the representation is got from LSTM RNN, the LSTM soft-max probability itself contains sequence training information. However, because of sequence training, the wrongly classified frames will gather together, as well as the right classification frames. It will lead to that the segment accuracy cannot improve so much as the frame accuracy when we use majority vote to get the segment labels and segment accuracy. To solve this problem, the LSTM segment feature and the initial segment feature are computed from the statistics of the LSTM frame features and the initial frame features respectively. Then the segment feature is used as the input to training the classification system and get the segment labels. Experimental results show that the LSTM segment feature can improve the classification accuracy by using segment feature for training and testing instead of majority voting. As the LSTM soft-max probability is tend to be linear and lost many information, we propose the fusional segment feature to further improve the performance. The

fusional feature combines the advantages of LSTM soft-max probability, and contains more information than LSTM segment feature. It is obtained by fusing the initial segment feature and the LSTM segment feature. Results show that the fusional segment feature can further improve the classification accuracy.

Erik Marchi [15] uses LSTM for identifying abnormal/novel acoustic signals. Besides this research, there have few studies of LSTM RNN in audio classification. To the best of our knowledge, this paper first uses LSTM for music genre classification.

2. LSTM RNN based Segment Features

2.1. Initial Feature Extraction

In this paper, we use the scattering feature as the initial feature. The scattering feature is an extension of the Mel-Frequency Cepstral Coefficient (MFCC). The MFCC feature is generally extracted by using small windows, whose typical duration is 25 ms. However, when the length of the window increases, the information lose will become significant. For music whose long time interval representation is helpful for classification, the MFCC has a poor classification performance. The scattering feature has been proved successful for music genres classification [16–18]. It builds invariant, stable and informative signal representations and is stable to deformations. It is computed by scattering the signal information along multiple paths, with a cascade of wavelet modulus operators implemented in a deep convolution network (CNN).

We extract the scattering feature using ScatNet [19], which is a toolbox for extracting scattering feature. In our work, we calculate first-order and second-order time scattering coefficients using a window of 370 ms with half overlap. The parameters for scattering transform are just the same as our previous work in [18].

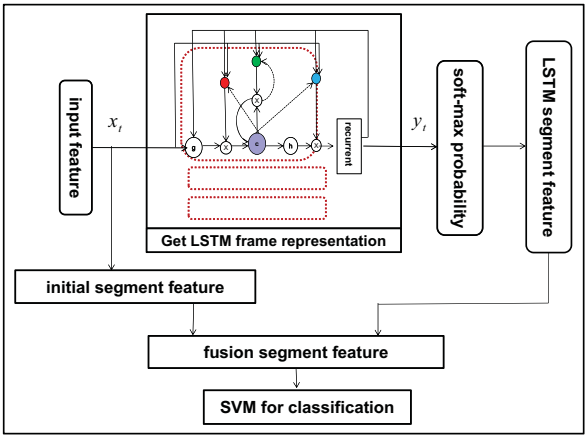

Figure 1: Architecture of proposed model.

2.2. Sequence Modeling using LSTM RNN

In a standard RNN, a input feature vector sequence is represented as $\mathbf{x} = (x_1, x_2, ..., x_T)$, then the hidden vector sequence $\mathbf{h} = (h_1, h_2, ..., h_T)$ and output vector sequence $\mathbf{y} = (y_1, y_2, ..., y_T)$ from $t = 1$ to T are computed using following equations:

$$h_t = \sigma(W^{xh}x_t + W^{hh}h_{t-1} + b^h) \quad (1)$$

$$y_t = W^{hy}h_t + b^y \quad (2)$$

Where $\sigma(\cdot)$ is activation function, the superscript 'x', 'h' and 'y' in 'W' and 'b' are represent the input layer, the hidden layer and the output layer respectively. For example, W^{xh} is the weight matrix connecting the input layer and the hidden layer, W^{hh} is the weight matrix connecting different hidden layers, b^h is the hidden bias vector, b^y is the output bias vector .

It is widely recognized in RNN that the activation function may led to the problem of vanishing gradient problem. To solve this problem, the hidden notes are replaced by a set of cells, which are called the LSTM memory blocks. Each LSTM memory cell contains three gates: input gate, output gate and forget gate. The forget gate is shown to be essential for problems with continual or very long input strings [12, 20]. In place of equations (1) and (2), the LSTM RNN is implemented according to the following equations [21]:

$$c_t = f_t \odot c_{t-1} + i_t \odot g(W^{cx}x_t + W^{cm}m_{t-1} + b^c) \quad (3)$$

$$m_t = o_t \odot h(c_t) \quad (4)$$

$$y_t = \phi(W^{ym}m_t + b^y) \quad (5)$$

The input gate i_t, forget gate f_t and output gate o_t are give by equations:

$$i_t = \sigma(W^{ix}x_t + W^{im}m_{t-1} + W^{ic}c_{t-1} + b^i) \quad (6)$$

$$f_t = \sigma(W^{fs}x_t + W^{fm}m_{t-1} + W^{fc}c_{t-1} + b^f) \quad (7)$$

$$o_t = \sigma(W^{ox}x_t + W^{om}m_{t-1} + W^{oc}c_t + b^o) \quad (8)$$

For recurrent projected layer, r_t is computed as:

$$r_t = W^{rm}m_t \quad (9)$$

Where $\odot$ is the element-wise product operator, $g(\cdot)$ and $h(\cdot)$ are the cell input and cell output activation functions, $\phi(\cdot)$ and $\sigma(\cdot)$ are other activation functions, c_t is the cell state vector, W (W^{ix}, W^{ic}, W^{fx}, W^{fc}, W^{ox}, W^{oc}) is weight matrix, b (b_c, b_y, b_i, b_f, b_o) is bias vector, and the superscript i, f, o, c, x, y represent input gate, forget gate, output gate, cell state, input layer and output layer respectively.

2.3. Segment Features based Model

As in Figure 1, we first train a LSTM RNN. Then we get the soft-max probability as the LSTM frame feature vector $\{\mathbf{p}_i|, i = 1, 2, ..., n\}$ (n is the number of frames in a music track). An initial segment feature f_{sc} is the mean of input frame scattering feature $\mathbf{x}_i$ over each segment (a segment is a music track):

$$f_{sc} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i \quad (10)$$

The LSTM segment feature f_{lstm} is computed from the statistics of LSTM frame feature [22]. It contains four parts: the maximum of LSTM frame feature in a segment (f_{lstm}^{max}), the Minimum of LSTM frame feature in a segment (f_{lstm}^{min}), the mean of LSTM frame feature in a segment (f_{lstm}^{mean}), and the percentage of frames which have probability higher than k (threshold value) in a segment (f_{lstm}^{k}). They can be computed as following:

$$f_{lstm}^{max} = \max\{\mathbf{p}_i|i = 1, ..., n\} \quad (11)$$

$$f_{lstm}^{min} = \min\{\mathbf{p}_i|i = 1, ..., n\} \quad (12)$$

$$f_{lstm}^{mean} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{p}_i \quad (13)$$

$$f_{lstm}^{k} = \frac{Count[|\mathbf{p}_i > k|]}{n} \quad (14)$$

The compute is done on each dimension of $\mathbf{p}_i$, and these features are represent by 6-dimensional matrix. The final fusional segment feature f_{fusion} is the fusion of f_{sc} and f_{lstm}, which is used as the input of the classification system to training and testing the classification system.

3. Experiment Setup

3.1. Data Prepare

The ISMIR database has been used as benchmarks for music genre classification by many researchers [8, 23–25]. The detail of the the databases is described as Table 1, and the training set an testing set have been defined by the database.

Table 1. Database Description

genre	tracks(train/test)	time duration(hours)
Classical	320/320	17.87/16.71
Electronic	115/114	10.48/9.97
Jazz/Blue	26/26	1.66/1.80
Metal/Punk	45/45	3.14/2.95
Rock/Pop	101/102	6.34/6.79
World	122/122	11.92/10.86
total	729/729	51.41/49.08

Before feature extraction, each audio file has been converted into a 22050Hz, 16 bit, and single channel WAV file. Then, we extract the scattering feature just as the section 2.1. After that, each frame of scattering feature is represented by a 525 dimensional vector.

3.2. Baseline Systems

The baseline systems use DNNs as classifier. The DNN used here is the Karel's DNN implementation in kaldi [26] with random initialization. It includes one input layer, some hidden layers and one output layer. Within each hidden layer, the sigmoid function is used as the active function. For output layer, the soft-max function is used to compute the posterior probability. The cross-entropy function is used as the objective function to optimize the DNN. No dropout function is used and the learning rate is $8 * 10^{-6}$. The training epochs is 100.

The frame scattering feature is used as the input to train the DNN. Then, we got the soft-max probability. Each predict frame label is determined by the genre which has the maximum probability. The majority voting is performed on the frame predict labels of testing sets to get the label of each music track. The segment accuracy is computed on the track labels. The result of baseline models are summarized in the Table 2, in which "Seg_acc" means the segment accuracy here.

Table 2. Classification result of baseline models and LSTM RNN based models using initial frame feature

model(layers)	Frame_acc	Seg_acc
baseline-DNN1(525-1024-6)	75.92%	85.32%
baseline-DNN2(525-1024-1024-6)	76.35%	84.35%
LSTM(525-512-6)	83.85%	87.52%
LSTM(525-1024-6)	83.78%	87.65%
LSTM(525-512-512-6)	81.66%	86.28%
LSTM(525-1024-1024-6)	80.44%	82.99%

4. Music Genre Classification

4.1. Experiment of LSTM RNN based model using initial frame feature

In this experiment, we use the frame scattering feature as the input to train the LSTM RNN. The LSTM RNN is implemented based on the Kaldi's nnet1 framework [21].In the testing stage, the frame scattering feature is used to get the soft-max probability output of LSTM RNN. We get the frame accuracy and segment accuracy just as the baseline models. The parameters are listed in Table 3. Other parameters use the default values.

Table 3. The parameters for LSTM RNN Training

parameter name	value	describe
learn_rate	0.00002	the learning rate
momentum	0.8	the momentum
batch_size	60	the batch size
max_iter	50	the training epochs
halving_factor	0.9	the attenuation rate of the learning rate
cellDim	800/1500	if the number of hidden nodes is 512/1024

The results of the baseline models and the LSTM based models are shown in Table 2. We can see that LSTM RNN based models outperform DNN based models. Also, we can see that the frame accuracy is improved more than the segment accuracy. It is also shown that the sequence training using LSTM RNN can perform better than independent frame training using DNN. But there is also a disadvantage in LSTM RNN training that the wrongly classified frames will gather together, as well as the right classification frames. It results that the segment accuracy cannot improve so much as the frame accuracy. The model using the segment features is proposed to solve this problem, and the experiment in next subsection will prove this.

4.2. Experiment of Segment Features based Model

In this experiment, we aim to get a further improvement by using the segment features. We first get LSTM frame feature. Then, we calculate f_{sc}, f_{lstm} and f_{fusion}, and each LSTM frame feature is represented by a 24-dimensional vector(combine 4 segment features which are described in equations 11-14). As each initial frame feature is a 525-dimensional vector, each fusional segment feature f_{fusion} will be represented by a 549-dimensional vector after computing. The classifier we used for classification is linear SVM. We respectively use LSTM segment feature f_{lstm} and fusion segment feature f_{fusion} as the input of SVM, and then compare the results.

Table 4. Classification results of segment features based model and other existing approaches

model(LSTM layer)	Seg_acc	LSeg_acc	FSeg_acc
LSTM(525-512-6)	87.52%	88.07%	**89.71%**
LSTM(525-1024-6)	87.65%	88.34%	88.07%
LSTM(525-512-512-6)	86.28%	86.14%	88.48%
LSTM(525-1024-1024-6)	82.99%	83.95%	86.42%
son2008 [27]	84.77%		
hol2008 [28]	83.5%		
lee2009 [23]	86.8%		
leo2012 [24]	76.27%		
sig2014 [8]	73.4%		

The results are shown in Table 4. In the table, "Seg_acc" is the segment accuracy using LSTM RNN to train the model and then perform the majority vote, "LSeg_acc" is the performance of model which uses segment feature f_{lstm} as the input of SVM, and "FSeg_acc" is the performance of model which uses segment feature f_{fusion} as the input of SVM. The k used for computing f_{lstm} and f_{fusion} here is 0.9. From Table 4, we can see that the LSTM segment feature f_{lstm} based model perform better than LSTM RNN based model using frame feature, and the model using fusional segment feature f_{fusion} leads a further improvement comparing with the model using LSTM segment feature f_{lstm}. Table 4 also lists other existing approaches for music genre classification on the ISMIR database.

4.3. The Analysis of Segment Features

In this section, we first analyze how the segment features based model improves the performance. Figure 2 shows the statistics of the LSTM soft-max probability score. In the Figure, the vertical black dotted lines divide each subgraph into 6 parts, and these six parts of score data are corresponding belongs to the first, second, third, forth, fifth, sixth category of music tracks. The probability score is a six dimensional vector, which each dimension represents the score of one music category.

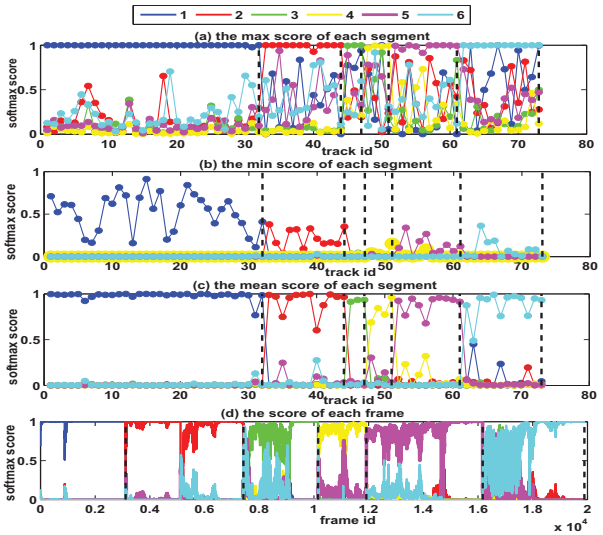

Figure 2: The subgraph a, b, c show the maximum, minimum and means soft-max score of six music categories in each segment (a music track). The subgraph d is soft-max probability score of each frame. The soft-max score in the figure is get from the model "LSTM(525-512-6)". The six categories of line (the legend "1", "2",...,"6") represent the score in six dimension, and each dimension represents the score of one music category. The"x-label" in the subgraph a, b, c (d) represent the segment index (frame index), and the "y-label" in four subgraphs is the corresponding maximum score (in six dimensions) of this segment (frame).

Figure 2(a) shows the maximum score of six music categories in each segment. We can see that the maximum score for different categories has a very clear distinction. For example, in the first part (for the first category of music), most segments have the highest score in the first dimension. The score of the first dimension means the score for the first category of music, and label of the highest score will be the predict label. It indicates that most segment in the first part will be right classified. This is the same as Figure 2(b) and 2(c). We can see that the maximum, minimum, average of a category of music will have an obviously higher score in corresponding dimension. Therefore these three parts of fusional segment feature are discriminative.

The forth part of fusional segment feature f_{lstm}^{k} is the percentage of frames which have probability higher than k in a segment. Figure 3 shows how k effects the performance. For example, the model "f_fusion(525-512-6)" has the best performance when $k = 0.95$. The choice of k depends on the soft-max score of LSTM. From Figure 2(d), we can see that if we use a higher k (e.g. $k >= 0.8$), the f_{lstm}^{k} will have a higher value in the dimension corresponding to its music category, and have a lower value in other categories. Then f_{lstm}^{k} will have a distinguishing difference among six categories of music tracks.

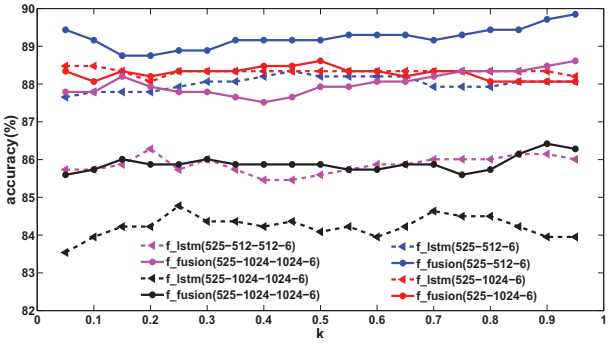

Figure 3: The performance using different k ('k' is described in equations 14). "f_lstm" means the model uses segment feature f_{lstm} as input feature and use linear SVM as classifier. "f_fusion" means the model uses fusional segment feature f_{fusion} as input feature and use linear SVM as classifier.

By comparing of the results in Table 2, Table 4, Figure 2 and Figure 2. It is obvious that LSTM RNN based models performs better than DNN based models for music genre classification. The statistics of LSTM frame feature is more discriminative than initial frame feature, and the proposed fusional segment feature further improves the classification accuracy. The proposed fusional segment feature model improves 4.19% classification accuracy compared to DNN based model, and improves 2.19% compared to LSTM RNN based model.

5. Conclusions

We propose the LSTM RNN based segment features for music genre classification. LSTM RNN is good at modeling feature sequence with long-term context, which can better represent the music data. But the sequence training of LSTM RNN will lead to that the wrongly classified frames gather together, as well as the right classification frames. Then the fusional segment feature which combines the LSTM RNN and scattering feature is proposed to ameliorate this problem. The significant improvement of proposed feature indicates the success of proposed model.

6. Acknowledgements

This research was supported by following two parts: The China National Nature Science Foundation (No. 61573357, No. 61503382No. 61403370, No. 61273267 and No. 91120303, No. 61305027); Technical development project of state grid corporation of China entitled "machine learning based Research and application of key technology for multi-media recognition and stream processing".

7. References

[1] A. Meng, P. Ahrendt, J. Larsen, and L. K. Hansen, "Temporal feature integration for music genre classification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 5, pp. 1654–1664, 2007.

[2] Y. Song and C. Zhang, "Content-based information fusion for semi-supervised music genre classification," *Multimedia, IEEE Transactions on*, vol. 10, no. 1, pp. 145–152, 2008.

[3] I. H. Chung, T. N. Sainath, B. Ramabhadran, M. Picheny, J. Gunnels, V. Austel, U. Chauhari, and B. Kingsbury, "Parallel deep neural network training for big data on blue gene/q," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 2015, pp. 745–753.

[4] Y. Zhao, D. P. Tao, S. Y. Zhang, and L. W. Jin, "Similar handwritten chinese character recognition based on deep neural networks with big data," *Journal on Communications*, 2014.

[5] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[6] I. Mcloughlin, H. Zhang, Z. Xie, Y. Song, and W. Xiao, "Robust sound event classification using deep neural networks," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 23, no. 3, pp. 540 – 552, 2015.

[7] X. Yang, Q. Chen, S. Zhou, and X. Wang, "Deep belief networks for automatic music genre classification," *ntM*, vol. 92, no. 11, pp. 2433–2436, 2011.

[8] S. Sigtia and S. Dixon, "Improved music feature learning with deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, 2014, pp. 6959–6963.

[9] H. Wang, L. Tan, and C. C. Leung, "Unsupervised spoken term detection with acoustic segment model," in *Speech Database and Assessments (Oriental COCOSDA), 2011 International Conference on*, 2011, pp. 106–111.

[10] H. Wang, C. C. Leung, T. Lee, B. Ma, and H. Li, "An acoustic segment modeling approach to query-by-example spoken term detection," *Gastroenterology*, vol. 142, no. 5, pp. S–205, 2012.

[11] T. Nakashika, T. Takiguchi, and y. Ariki, "High-order sequence modeling using speaker-dependent recurrent temporal restricted boltzmann machines for voice conversion," in *Interspeech*, 2014.

[12] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, 2015.

[13] D. Renshaw and K. B. Hall, "Long short-term memory language models with additive morphological features for automatic speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, 2015.

[14] G. Chen, C. Parada, and T. N. Sainath, "Query-by-example keyword spotting using long short-term memory networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, 2015.

[15] E. Marchi, F. Vesperini, F. Eyben, and S. Squartini, "A novel approach for automatic acoustic novelty detection using a denoising autoencoder with bidirectional lstm neural networks," in *Proceedings - ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015.

[16] J. Andén and S. Mallat, "Deep scattering spectrum," *CoRR*, vol. abs/1304.6763, 2013.

[17] J. Anden and S. Mallat, "Multiscale scattering for audio classification," in *ISMIR*, 2011, pp. 657–662.

[18] J. Dai, W. Liu, C. Ni, L. Dong, and H. Yang, "Multilingual deep neural network for music genre classification," in *Interspeech*, 2015.

[19] "scattering," http://www.di.ens.fr/data/scattering/.

[20] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: continual prediction with lstm," *Neural Computation*, vol. 12, no. 10, pp. 2451–2471, 2000.

[21] "The implemente of lstm," https://github.com/dophist/kaldi-lstm.

[22] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Interspeech 2014*, September 2014. [Online]. Available: http://research.microsoft.com/apps/pubs/default.aspx?id=230136

[23] 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," *Multimedia, IEEE Transactions on*, vol. 11, no. 4, pp. 670–682, 2009.

[24] F. D. Leon and K. Martinez, "towards efficient music genre classification using fastmap," *Proc Dafx*, 2012.

[25] Ismir 2004 genre dataset. [Online]. Available: http://ismir2004.ismir.net/genre_contest/index.html

[26] "Kaldi," http://kaldi.sourceforge.net.

[27] Y. Song and C. Zhang, "Content-based information fusion for semi-supervised music genre classification," *Multimedia IEEE Transactions on*, vol. 10, no. 1, pp. 145–152, 2008.

[28] A. Holzapfel and Y. Stylianou, "Musical genre classification using nonnegative matrix factorization-based features," *IEEE Transactions on Audio Speech and Language Processing*, vol. 16, no. 2, pp. 424–434, 2008.