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

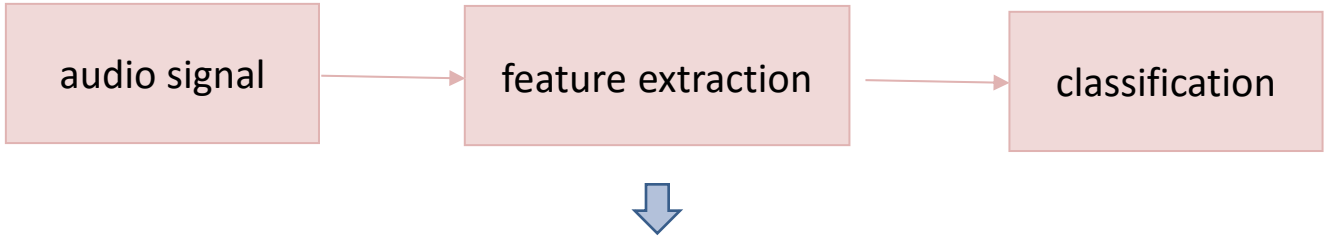
----for ISCSLP 2016

Jia Dai, Shan Liang, Wei Xue, Chongjia Ni, Wenju Liu

National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, Beijing, China, 100190

Introduction

Audio classification system:



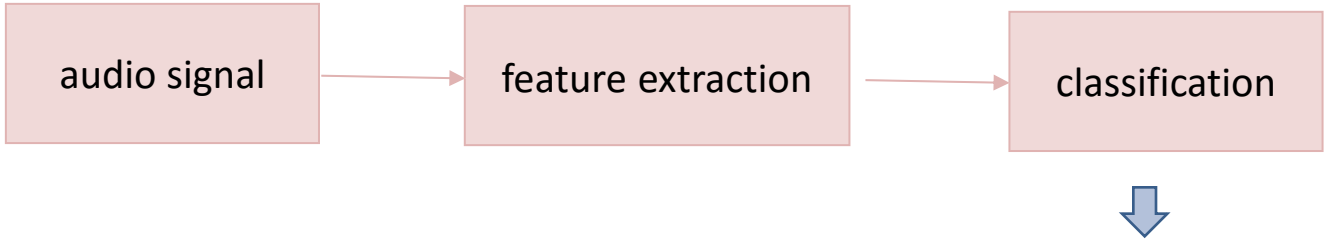
feature model

aims at finding more discriminative features

- **base features**
  MFCC, LPCC, Fbank.....
  silence ratio, High Zero Crossoing Rate Ration.....
- **the combination of different features**
- **the transform of different features**
  sparse, Bag-of-feature, DNN for feature learning....

Introduction

Audio classification system:



classification model

focuses on designing better classification models to better capture the discrimination of features belonging to different classes

- **single classifier**
  KNN, SVM, GMM, DNN, RNN,CNN.....
- **score fusion of different classifiers**
  strong classifier = w1*weak classifier 1+ w2* weak classifier2+,.....
- **two stages classification**
  GMM-DNN, DNN-SVM, DNN-DNN, LSTM-SVM...

Introduction

Background:

- 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.
- 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.
- The LSTM RNN has been successfully used for many sequence labeling and sequence prediction tasks.

Introduction

Background:

- 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-RNN based segment features are proposed.

Feature

MFCC

- MFCC is effective over the length of a short time window 25 ms, and when using a window larger than 25 ms, the information lose becomes too important [1]. For music whose long time interval representation helps more for classification, the MFCC have a poor classification performance.
- Scattering transform is an extension of MFCC. It has been proved successful for music genres classification [1] [2].

[1] J. And´en and S. Mallat, "Deep scattering spectrum," CoRR, vol. abs/1304.6763, 2013
[2] J. Anden and S. Mallat, "Multiscale scattering for audio classification," in ISMIR, 2011, pp. 657–662.

Feature

- The scattering feature.
  - Scattering feature is computed by scattering the signal information along multiple paths, with a cascade of wavelet modulus operators implemented in a deep convolution network (CNN) [1].
  - In the Figure. 1 the first order and second order of the scattering transform are shown, which is much like the structure of CNN. And it can be expanded to the third order or more.

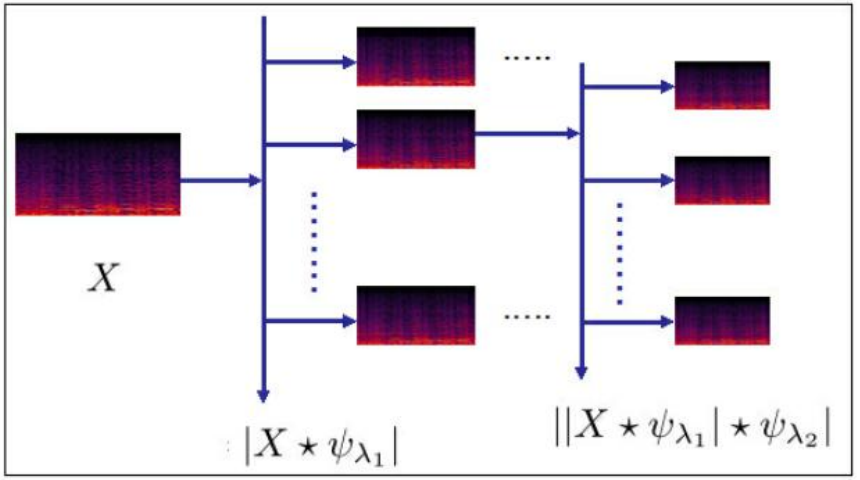


Figure 1: The first order and second order scattering transform

[1] J. And´en and S. Mallat, “Deep scattering spectrum,” CoRR, vol. abs/1304.6763, 2013

Data

The ismir dataset:

- The ISMIR database is used to train the system and test the performance.
- Before feature extraction, each audio file in ISMIR has been converted into a 22050Hz, 16 bit, and single channel WAV file.

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

Framework

- LSTM RNN based Segment Features is proposed . It contains three main stages.

  - Initial Feature Extraction: the scattering feature
  - Sequence Modeling using LSTM RNN
  - Segment Features based Model

Framework

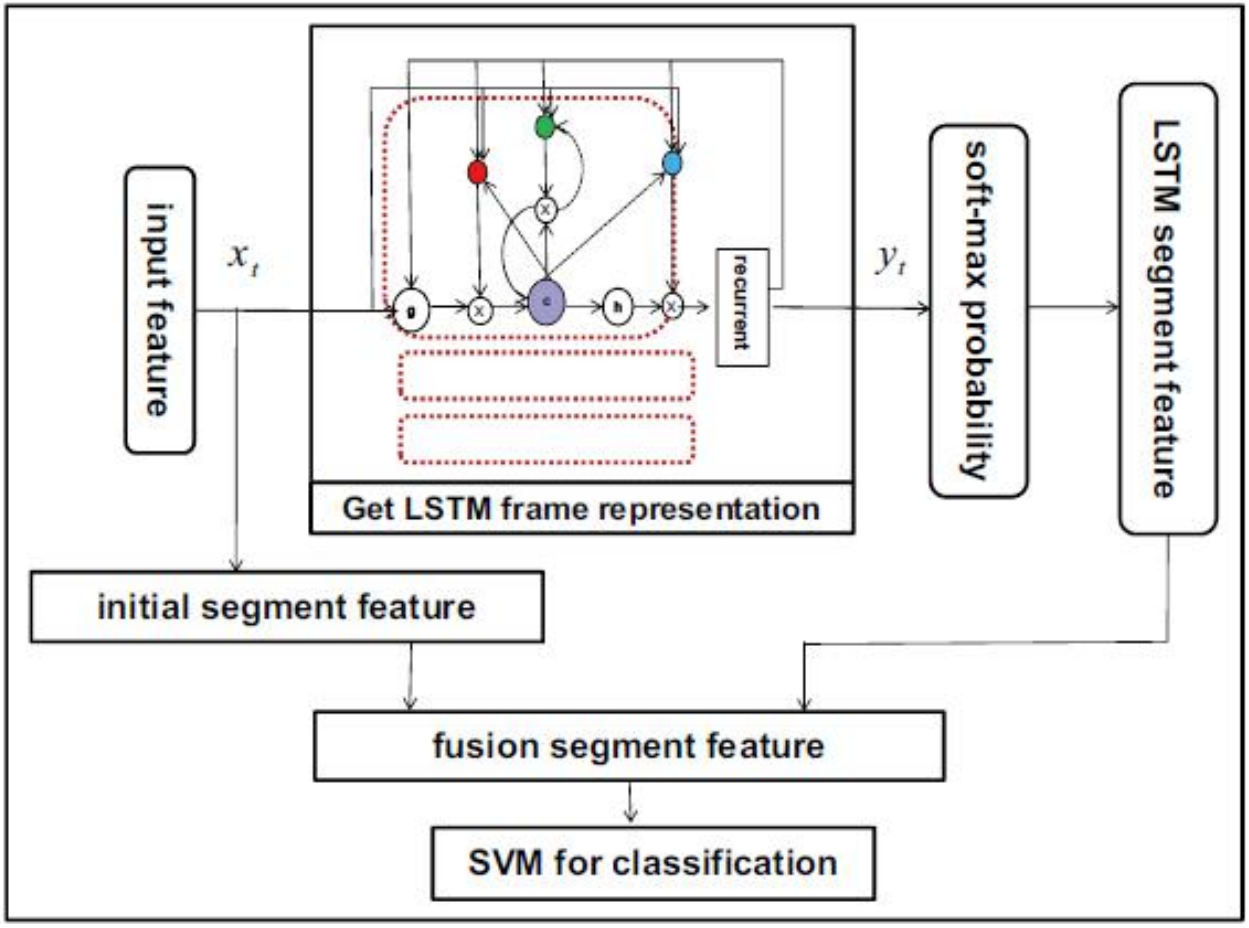


Figure 2.The architecture of proposed model

Framework

- Initial Feature Extraction: the scattering feature

  - We extract the scattering feature using ScatNet [3], which is a toolbox for extracting scattering feature.
  - The scattering feature is used as input feature and initial frame feature. Then we use the mean of scattering feature in a track as the initial segment feature as in Figure 2.

[3] “scattering,” http://www.di.ens.fr/data/scattering/.

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

LSTM-RNN based System

- Sequence Modeling using LSTM RNN

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

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

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

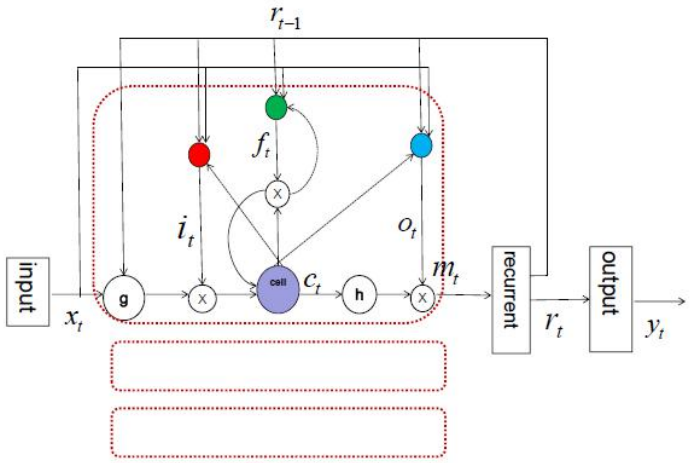
- To the problem of vanishing gradien, the hidden notes are replaced by a set of cells, which are called the LSTM memory blocks.

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

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

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

- Each LSTM cell contains a cell and three gates: input gate, output gate and forget gate.



Analysis of LSTM-RNN based system

- That the sequence training using LSTM RNN can perform better than independent frame training using DNN.
- But we can see that the frame accuracy is improved so much, while segment accuracy is improved so litter. why?
- 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.

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%

Segment Features based Model

- The LSTM segment feature is computed from the statistics of LSTM frame feature . It contains four parts:

  - the maximum of LSTM frame feature in a segment
  - the Minimum of LSTM frame feature in a segment
  - the mean of LSTM frame feature in a segment
  - the percentage of frames which have probability higher than k (threshold value) in a segment

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

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

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

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

Analysis of proposed system

- The segment features have a distinguishing difference among six categories of music tracks. Therefore segment features are more discriminative than frame feature.
- Figure 3(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.

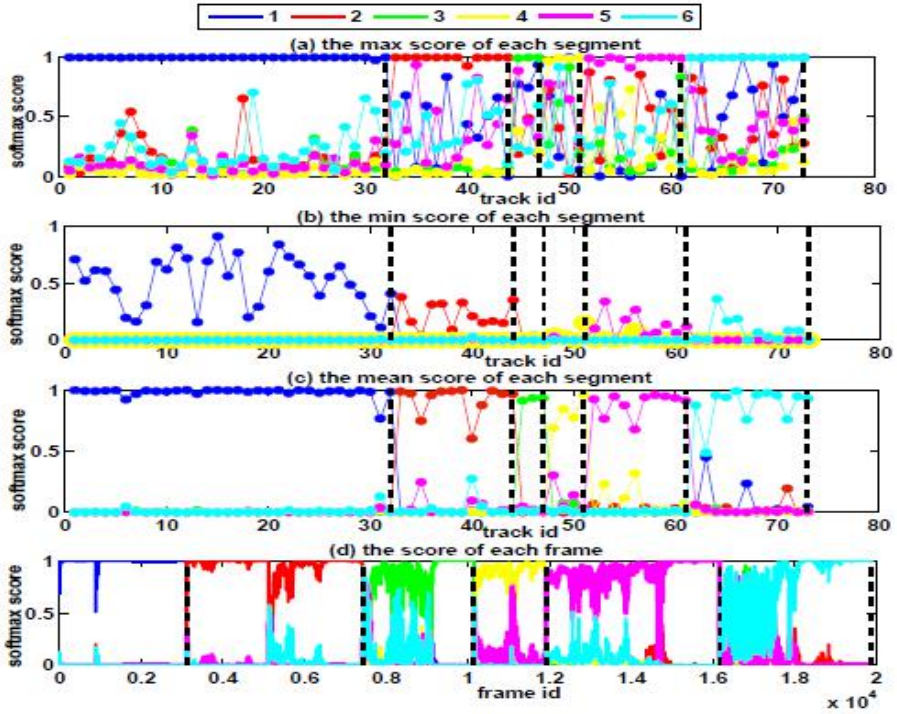


Figure 3: The subgraph a, b, c show the maximum, minimum and means soft-max score of six music categories in each segment (a music track). The subgraph d is soft-max probability score of each frame. The soft-max score in the figure is get from the model "LSTM(525-512-6)". The six categories of line (the 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).

Experiment setup

setup:

- The baseline-DNN we used here is Karel's DNN implementation in kaldi with random initialization.
- No dropout function is used in baseline-DNN.
- The learning rate used in baseline-DNN is 8*10^(-6) .
- The training step of baseline-DNN is 100.
- The learning rate used in LSTM-RNN is 0.00002 .
- The momentum used in LSTM-RNN is 0.8 .
- The batch size used in LSTM-RNN is 60 .
- The training step of LSTM-RNN is 100.
- The the attenuation rate of the learning rate in LSTM-RNN is 0.9.

Results and Conclusions

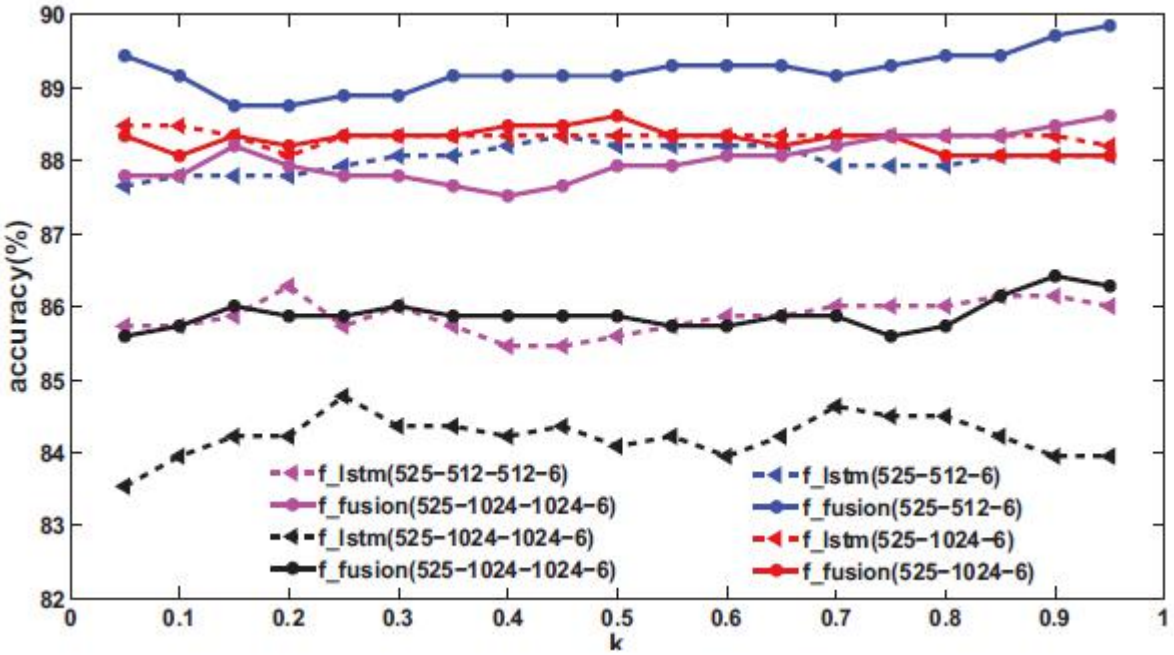
- The proposed fusional segment feature model improves 4.19% classification accuracy compared to DNN based model, and improves 2.19% compared to LSTM RNN based model.

Table 3. 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%		

Results and Conclusions

- The segment features haIt is obvious that LSTM RNN based models performs better than DNN based models.
- The statistics of LSTM frame feature is more discriminative than initial frame feature, and the proposed fusional segment feature further improves the classification accuracy.




NLPR
LABORATORY OF PATTERN RECOGNITION
NATIONAL


CHINESE ACADEMY OF SCIENCES

Thanks!