THINKIT AUDIO GENRE CLASSIFICATION SYSTEM FOR MIREX08

Chuan Cao, Ming Li
ThinkIT Speech Lab., Institute of Acoustics,
Chinese Academy of Sciences,
{ccao,mli}@hccl.ioa.ac.cn

ABSTRACT

This full abstract describes our submitted system for the MIREX08 Audio Genre Classification task, the goal of which is to discriminate music excerpts of different genres/styles. The system is based on basic feature of MFCC and modeling framework of GSV-SVM, which has been successfully applied in speaker recognition field. In this submission, the only basic feature we use is MFCC. And the goal of this submission is to test the performance of pure GSV-SVM framework on music genre classification task.

1 INTRODUCTION

Music genres are defined as labels of different music styles and used for categorizing the vast amount of music from all the world. The huge amount has brought great challenge to the organization and retrieval of music databases and also brought great opportunities to automatic music information retrieval (MIR) techniques. Numbers of genre classification methods have emerged in last few years, such as Marsyas [7] and G1C [4]. The various systems use different features and different modeling frameworks and the Music Information Retrieval Evaluation eXchange (MIREX) 2008 Audio Genre Classification task provides a common platform to compare and evaluate these state-of-the-art music genre classification algorithms [1].

The system introduced in this abstract is based on a Gaussian Super Vector followed by Support Vector Machine (GSV-SVM) framework, which has been successfully applied in speaker recognition field. And only short-time spectral features are used to test the performance of pure GSV-SVM framework. Specifically, Mel Frequency Cepstral Coefficients (MFCC) are extracted from every music excerpt and super vector is obtained by the mapped Gaussian mixture model (GMM), which is mapped from a universe background model (UBM) with all the MFCCs of the excerpt. Then the super vector is seemed as feature and support vector machine is utilized to train models of different genres. Similarly, genre labels of test music pieces are decided by SVM classification result on the super vector feature of the test music.

2 BASIC FEATURES

2.1 Mel Frequency Cepstral Coefficients

Mel-frequency cepstral coefficients (MFCC) are a set of perceptual motivated features that describe the shape of the short-time Fourier spectrum. MFCC has been widely used for short-time spectral modeling and successfully applied in speech recognition society since 1980s.

In our system, 36 dimensions MFCC features (18 base coefficients and their delta among adjacent frames) are extracted for every 10ms frame. All the training and modeling processes backward are based only on this set of features. No other temporal features or long-time features is used, since the goal of this submission is to test the performance of pure GSV-SVM framework on music genre classification task.

3 GSV-SVM FRAMEWORK

GSV-SVM is the abbreviation of Gaussian super vector followed by support vector machine, which was proposed by Campbell in 2006 [2] and has been successfully applied in speaker recognition field. A high-dimensional super vector is mapped from original features (e.g. MFCCs), by the algorithm of maximum a posterior (MAP) from a pre-trained universal background model (UBM) [5]. Then training and modeling processes are all based on the super vector feature, by a SVM classifier tool.

3.1 Universal Background Model

The universal background model (UBM) is a large Gaussian mixture model (GMM) trained to represent the common distribution of input features. In speaker recognition framework, UBM is usually used to model speaker-independent distribution and is trained on a large amount of reflective and balanced speech utterances.

In our submission for music genre classification, the UBM is trained from two datasets. One is the GTZAN dataset, which was originally compiled by George Tzanetakis [6] and consists of 1000 music excerpts equally distributed over 10 popular music genres. Another is a dataset collected by ourselves, which consists of nearly 2000 audio pieces over 9 genres.

3.2 MAP and Super Vector

MAP algorithm is a widely used technique in speaker recognition systems, such as GMM-UBM system and GSV-SVM system. In MAP framework, parameters to be estimated can be seen as a weighted sum of old parameters (UBM parameters) and new parameters (parameters derived from observation). Given a pre-trained UBM and the observations, the adapted parameters (only the means are adapted in this method) can be calculated as:

$$\mu_i^* = \alpha^m E_i(x) + (1 - \alpha^m)\mu_i \tag{1}$$

where α^m refers to the weighting factor and μ represents parameters of UBM. $E_i(x)$ is obtained by:

$$E_i(x) = \frac{\sum_{t=1}^{T} P_r(i|x_t)x_t}{\sum_{t=1}^{T} P_r(i|x_t)} \tag{2}$$

in which, $P_r(i|x_t)$ is the posterior of mixture i given by x_t

$$P_r(i|x_t) = \frac{w_i p_i(x_t)}{\sum_{i=1}^{M} w_j p_j(x_t)} \tag{3}$$

where w_i is the weight of the i-th Gaussian of UBM and $p_i(x_t)$ is the likelihood of x_t on the i-th Gaussian. Therefore, we can get an adapted GMM with new estimated means. The means of the adapted GMM are normalized by corresponding variances and weights and then concatenated as a super vector which will be modeled and classified by SVM back-end.

3.3 SVM Training and Testing

Support vector machine (SVM) is a very powerful two-class classifier that has been widely and successfully used in nearly every corner of pattern recognition field. In our submission, we use the SVM-Light toolkit released by Cornell [3] as the SVM back-end. Specifically, for every genre, the SVM model is trained from pieces of the target genre (as positive samples) and pieces of all other genres (as negative samples). For testing, every test piece is scored over all the genre models and then it is assigned with the label of the maximum score.

4 IMPLEMENTATION

The algorithm is implemented in C++ and is for Windows platform. The execution time is nearly two hours on MIXED dataset and one hour on LATIN dataset, including all the procedures of feature extraction, model training and classification. The UBM is pre-trained by ourselves and included in the submission package and the MAP procedure is doing out of the feature extraction process.

Team	MIXED dataset	LATIN dataset
CL1	62.04%	65.17%
CL2	63.39%	64.04%
GP1	63.90%	62.72%
GT1	64.71%	53.65%
GT2	66.41%	53.79%
GT3	65.62%	53.67%
LRPPI1	65.06%	58.64%
LRPPI2	62.26%	62.23%
LRPPI3	60.84%	59.55%
LRPPI4	60.46%	59.00%
ME1	65.41%	54.15%
ME2	65.30%	54.70%
ME3	65.20%	54.99%

Table 1. The evaluation results of MIREX08 Audio Genre Classification task, arranged in alphabetical order. See [1] for more details.

5 EVALUATION RESULTS AND DISCUSSION

Two systems (CL1 and CL2) are submitted for the genre classification task. The only difference between them is the number of Gaussian mixture. The evaluation results are listed in Table.1.

As can be seen in Table.1, our systems can correctly classify 63.39% of all the test data in MIXED dataset, while the best system reaches the accuracy of 66.41% (GT2). While for LATIN dataset, our submissions perform the best among all the participants. Specifically, CL1 reaches 65.17% classification accuracy and CL2 got the accuracy of 64.04%. We are now making efforts to find out the reason for the large performance difference. But we also can see that the raw classification accuracy of our systems are very consistent on both dataset (nearly 64% accuracy), while most of other submissions are biased on the MIXED dataset. We think it may due to the reason that we use only simple short-time spectral features (MFCC) and they are quite robust in most situations. While other submissions utilize other more complicated features and parameters, which may not work or need to be tuned specifically for LATIN music, since this new dataset is introduced in by this year. And this maybe one of the reasons to explain the results.

6 ACKNOWLEDGEMENTS

Many thanks to the IMIRSEL team at the University of Illinois at Urbana-Champaign (UIUC) for organizing and running the MIREX evaluations, which has supplied an great information exchange platform for the MIR community.

References

[1] *http://www.music-ir.org/mirex/2008/index.php*.

[2] WM Campbell, JP Campbell, DA Reynolds, E. Singer, and PA Torres-Carrasquillo. Support vector machines for speaker and language recognition. *Computer Speech & Language*, 20(2-3):210–229, 2006.

[3] T. Joachims. SVM light support vector machine [EB/OL]. *URL: http://svmlight.joachims.org*, 2002.

[4] E. Pampalk. Computational Models of Music Similarity and their Application in Music Information Retrieval. *Docteral dissertation, Vienna University of Technology, Austria, March*, 2006.

[5] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn. Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, 10(1-3):19–41, 2000.

[6] G. Tzanetakis. Manipulation, Analysis AND Retrieval Systems FOR Audio Signals. *Department Of Computer Science*.

[7] G. Tzanetakis and P. Cook. MARSYAS: a framework for audio analysis. *Organised Sound*, 4(03):169–175, 2000.