THINKIT'S SUBMISSIONS FOR MIREX2009 AUDIO MUSIC CLASSIFICATION AND SIMILARITY TASKS

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 systems for the MIREX09 audio classification tasks (genre, mood, classical composer, audio tagging) and music similarity and retrieval task. All the classification systems are based on basic acoustic features (e.g. MFCC) and the modeling framework of GSV-SVM, which has been successfully applied in speaker recognition field. And the similarity systems are based on a simple Euclid distance measurement on the mid-level features, which are also mapped from the basic acoustic features.

1. INTRODUCTION

The huge amount of music has brought great challenge to the organization and retrieval of music databases and also brought great opportunities to many music information retrieval (MIR) techniques, such as automatic music genre/mood/artist classification and audio music similarity measurement. Although genre/mood/artist are notions from three different prospect, they all describe the similarity between music audios on a certain extent. Normally, the basic ideas and framework of genre/mood/artist classification systems are much the same, so we take the genre system as an example to describe afterward.

Numbers of genre classification methods have emerged in last few years, such as Marsyas [1] and G1C [2]. The various systems use different features and different modeling frameworks and the Music Information Retrieval Evaluation eXchange (MIREX) 2009 provides a common platform to compare and evaluate state-of-the-art music genre classification algorithms [**?**].

The classification system introduced in this abstract is based on a Gaussian Super Vector followed by Support Vector Machine (GSV-SVM) framework. A set of acoustic features such as MFCC, rhythm pattern (RP) are extracted from every music excerpt in the front-end part. To be different from G1 method, frame-level features are used to adapt a Gaussian mixture model (GMM) from a music universe background model (UBM) in our system. Then a super vector representing the adapted GMM model is obtained and 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. GSV-SVM FRAMEWORK

GSV-SVM is the abbreviation of Gaussian super vector followed by support vector machine, which was proposed by Campbell in 2006 [3] and has been successfully applied in speaker recognition field. A high-dimensional super vector is mapped from low-level features (e.g. MFCCs), by the algorithm of maximum a posterior (MAP) from a pre-trained music universal background model (UBM) [4]. Then training and modeling processes are all based on the super vector feature, by a SVM classifier tool.

2.1 Universal Background Model

The music universal background model (UBM) is a large Gaussian mixture model (GMM) trained to represent the common distribution of low-level features. In our submission for music classification, the UBM is trained from a large amount of music audios collected by ourselves. The dataset consists nearly 2000 audio pieces over different genres.

2.2 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

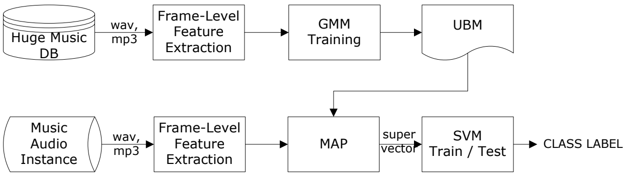

Figure 1. GSV-SVM framework.

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.

2.3 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 [5] 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.

3. ABOUT THE TAGGING AND SIMILARITY SYSTEMS

The submitted tagging systems are much like the train-test systems, since there is a train/test procedure too. The only difference is every sample of tagging task have multiple tag labels. We consider all the samples without a tag as the negative samples for training. The affinity of a test audio over a specific tag is the classifying score on the target tag model.

For the similarity system, we submitted two systems. One measures the audio similarity by the distance of the Gaussian super vector in the Euclidean space. And the other measure the distance on a high-level. We use a set of pre-trained genre models to map the Gaussian super vector features to a set of classification scores over the genre models. Then the Euclid distance of the genre scores vector is used for similarity measurement.

Team	Genre(Mixed)	Genre(Latin)	Mood	Composer
ANO	60.50%	38.87%	50.67%	41.77%
BP1	70.63%	67.31%	58.17%	55.66%
BP2	68.51%	63.52%	59.67%	54.76%
CL1	73.23%	**74.66**%	**65.67**%	60.97%
CL2	**73.33**%	73.58%	65.50%	60.03%
FCY1	–	–	60.33%	–
FCY2	–	–	58.33%	–
GLR1	71.23%	62.79%	60.83%	55.34%
GLR2	60.14%	49.84%	53.00%	45.92%
GP	64.24%	62.63%	63.67%	48.85%
GT1	65.10%	49.75%	59.33%	43.69%
GT2	67.87%	52.82%	56.83%	51.48%
HNOS1	64.47%	56.32%	58.67%	43.33%
HNOS2	20.90%	–	34.67%	15.84%
HNOS3	64.34%	56.22%	58.67%	42.24%
HNOS4	45.16%	30.05%	51.17%	29.04%
HW1	65.99%	54.72%	61.33%	56.35%
HW2	65.31%	52.28%	61.67%	53.10%
LZG	68.29%	55.96%	61.67%	54.40%
MTG1	64.79%	61.68%	57.67%	54.73%
MTG2	–	62.39%	57.50%	**62.05**%
MTG3	64.06%	45.80%	59.83%	48.12%
MTG4	64.00%	47.79%	59.33%	48.20%
MTG5	70.44%	61.14%	62.83%	49.75%
MTG6	–	63.16%	59.50%	50.36%
RCJ1	32.50%	38.93%	–	–
RCJ2	–	52.43%	–	–
RCJ3	37.71%	46.78%	–	–
RCJ4	50.99%	55.22%	–	–
RK1	61.41%	57.11%	53.17%	48.41%
RK2	–	–	41.33%	–
SS	66.60%	64.69%	58.83%	52.56%
TTOS	67.89%	53.70%	56.83%	44.37%
VA1	68.84%	58.37%	59.33%	53.57%
VA2	67.39%	54.49%	60.17%	53.57%
XLZZG	68.93%	55.25%	57.00%	53.54%
XZZ	69.36%	55.25%	60.00%	57.18%

Table 1. The evaluation results of MIREX09 train-test tasks, arranged in alphabetical order. See [6] for more details.

4. IMPLEMENTATION

The algorithm is implemented in C++ and is for Windows platform. The execution time is very quite short according to last year's statistics, including all the procedures of feature extraction, model training and classification. The UBM is pre-trained by ourselves and the MAP procedure is doing out of the feature extraction process.

5. EVALUATION RESULTS AND DISCUSSION

Two systems (CL1 and CL2) are submitted for all the train-test tasks. Besides, four tagging systems (CL1 CL4) and two similarity systems (CL1, CL2) are submitted. The train-test evaluation results are listed in Table.1. The result of similarity and retrieval results are shown in Table.2. Since there are lots of statistics in Audio Tagging tasks, we do not list the tagging result here.

As can be seen in Table.1, our systems have the best performance among all the participants on all the train-test tasks except the classical composer identification task, on which our submissions performed the 2^{nd} best. The results show the robustness and effectiveness of our systems,

Team	Average FINE Score	Average BROAD Score
ANO	5.391	1.126
BF1	2.401	0.416
BF2	2.587	0.410
BSWH1	5.137	1.094
BSWH2	5.734	1.232
CL1	2.525	0.476
CL2	5.392	1.164
GT	5.343	1.126
LR	5.470	1.148
ME1	2.331	0.356
ME2	2.585	0.418
PS1	5.751	1.262
PS2	**6.458**	**1.448**
SH1	5.042	1.012
SH2	4.932	1.040

Table 2. The evaluation results of MIREX09 Audio Music Similarity and Retrieval task, arranged in alphabetical order. See [6] for more details.

which are based on the GSV-SVM framework with set of low-level acoustic features. We have noticed that our advantage on the Genre(Latin) is much larger than the other three. We think maybe other participants have used some features effective for western music classification and those features failed to be robust on Latin music, which is much less concerned by the MIR researchers. And our systems use only raw simple and robust acoustic features that maybe fit well to Latin music too. Also we noticed that on the classical composer identification task, our systems failed to be the best. We attribute it to the reason that composing style is a higher level information and relatively more independent of low-level acoustic features. Since our systems do not use any high-level (music content level) features, we may fail to grasp information sensitive to composing styles. In our future work, music content features will be concerned and merged into the current systems aiming to represent high-level information.

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.

7. REFERENCES

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

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

[3] 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.

[4] 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.

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

[6] *http://www.music-ir.org/mirex/2009/index.php/MIREX2009_Results*.