A GENERIC TRAINING AND CLASSIFICATION SYSTEM FOR MIREX08 CLASSIFICATION TASKS: AUDIO MUSIC MOOD, AUDIO GENRE, AUDIO ARTIST AND AUDIO TAG

Geoffroy Peeters
Ircam Sound Analysis/ Synthesis Team - CNRS STMS
1, pl. Igor Stravinsky - 75004 Paris - France
peeters@ircam.fr

ABSTRACT

This extended abstract details a submission to the Music Information Retrieval Evaluation eXchange (MIREX) 2008 for the training and classification tasks audio music mood, audio genre, audio artist and audio tag. The same system has been submitted for the various tasks without any adaptations to the specific problems. The system named ircamclassification is a generic system which performs batch feature extraction, models training (using various classifiers) and file indexing (or file segmentation) into classes. The features extracted are generic in order to be applicable to many different audio and music indexing problems. The features are not specific to the above mentioned MIREX08 tasks. The goal of this submission is to test the applicability of a generic classification system to those tasks.

1 SYSTEM DESCRIPTION

ircamclassification is an extension of a system initially developed for instrument-samples indexing described in [3] using the features described in [4]. Only the subset of features applicable to polyphonic audio signals (music) has been used here. In [5] the system has been extended for speech/music segmentation. It is this system that has been used for MIREX08 tasks. We briefly review it in the following.

2 FEATURE EXTRACTION

In the present submission, only three sets of audio features are extracted from the signal.

MFCC: The first set aims at describing the shape of the spectrum at each time. Mel Frequency Cepstral Coefficients (40 Mel bands, 13 coefficients including DC component) are extracted every 20ms using a Blackman window of length 40ms.

SFM/ SCM: MFCCs only describes the shape of the spectrum whatever the content of the signal is noise or sinusoidal (harmonic) components. In order to describe this noise/ sinusoidal content, we also compute height Spectral Flatness [2] and Spectral Crest Measure coefficients. This is done using the same analysis parameters.

Chroma/ PCP: The third set of features gives rough information about the meaning of the harmonic content of the signal. For this, twelve Chroma [6]/ Pitch Class Profiles (PCP) [1] coefficients are computed using a Blackman window of length 100ms synchronized in time with the two other feature sets.

Delta and acceleration coefficients of the above mentioned features are also computed.

Finally, a simple temporal modelling (mean and standard deviation) of each feature is performed using a sliding window of length 500ms and a hop size of 250ms.

3 MODELS TRAINING

Training of the class-models is performed using the following steps:

Feature processing: Features are first normalized and outliers are removed (based on IQR).

Feature selection: The Inertia Ratio Maximization with Feature Space Projection (IRMFSP) algorithm [3] is used to select independently the best 40 features (independently means that we don't take into account the set the features belong to).

Feature space transform: Linear Discriminant Analysis is then applied to the reduced feature space.

Class modelling

Class modelling is done in two stages

First stage: frame-statistical-model We first model the belonging of each frame to each class using a simple Gaussian Mixture Models (8 Gaussians, full matrix). For this we use all the feature vectors $\underline{f}(t)$ for all the time $t \in J_k$ where J_k is the set of tracks labelled as belonging to class k. We call this model a frame-statistical model: it gives the probability to observe class k given the feature vector at time t: $p(t \in c_k|\underline{f}(t))$. As explained in [5], the labels are assigned to the tracks (a collection of frames) and not independently to the frames. A track of a given class may in fact include frames from another class: a track labelled as rock may contain frames belonging to the

blues class. It is the succession of the frame-belongings that makes the track being rock. We model this in the second stage of the classifier.

Second stage: track-statistical-model In the second stage we model the probability that the whole track belong to a class given the set of probability-vectors of its frame: $p(J \in c_k|\underline{p}(t_J \in c)$, where J is a track, t_J is the set of frames belonging to track J and $\underline{p}(t \in c)$ is the probability-vector coming from the frame-statistical model. For this, the whole training set is first classified using the frame-statistical-model. For each track belonging to class c_k we then study the belonging of its frames over time. This allows creating a track-statistical model.

4 CLASSIFICATION

The classification of an unknown track is also performed in two stages:

- first at the frame level using $p(t \in c_k|\underline{f}(t))$,
- then at the track level using $p(J \in c_k|\underline{p}(t_J \in c)$.

The training and classification process is represented in Figure 1.

5 RESULTS

The same system was submitted for all the tasks (named GP1 in the following). The system was initially only submitted for the Mood and Genre tasks. The IMIRSEL team asked us for the authorization to also run the system on the Artist and Tag tasks which we accepted. However, since the system was not supposed to deal with unbalanced sets or large amount of classes, some bugs appeared while running the system. We then provided a simplified version of the system based on a simple Gaussian modelling of the classes (named GP2 in the following) instead of a GMM modelling (GP1).

Since the system relies on automatic feature selection (IRMFSP algorithm) it is difficult to comment on the impact of the use of our specific features on the obtained results; we simply don't know which features the system has chosen!

5.1 Audio Music Mood Classification

The system submitted by Ircam (GP1) ranked first for the Audio Music Mood task (see Table 1).

5.2 Audio Mixed and Latino Genre Classification

The system submitted by Ircam ranked 5th (team ranking) for the Mixed Genre task and 2nd (team ranking) for the Latino Genre task (see Table 2).

Audio Music Mood Classification	
Participant	Average Classifcation Accuracy
GP1	**63.67%**
GT3	58.20%
LRPPI1	56.00%
LRPPI4	55.50%
LRPPI2	55.50%
GT1	55.00%
LRPPI3	54.50%
GT2	52.50%
ME1	50.33%
ME2	50.00%
KL	49.83%
ME3	49.67%
HW	30.33%

Table 1. Audio Music Mood Classification Results

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

Figure 2. Audio Mixed and Latino Genre Classification Results

Audio Artist Identification		Classical Composer Identification	
Participant	Average Classifcation Accuracy	Participant	Average Classifcation Accuracy
ME1	47.65%	ME1	53.25%
ME3	47.25%	ME2	53.10%
ME2	47.16%	ME3	52.89%
GT2	43.47%	**GP1**	**48.99%**
LRPPI1	35.42%	GT2	45.82%
GT3	35.27%	GT3	43.81%
GT1	33.66%	LRPPI4	39.54%
LRPPI2	33.20%	GT1	39.47%
LRPPI4	32.52%	LRPPI2	39.43%
LRPPI3	29.87%	LRPPI3	37.48%
GP2	**1.11%**	LRPPI1	34.13%

Table 2. Audio Artist and Classical Composer Identification Results

5.3 Audio Artist and Classical Composer Identification

The system submitted by Ircam ranked 2nd (team ranking) for the Audio Classical Composer Identification task and last for the Audio Artist Identification task (see Table 2).

One of the main characteristics of the Audio Artist Identification task is to have a large number of classes (105) to be trained with a small number of training examples (20 segments of 30s length for each class). Our system is a two-stage classifier. The inputs of the second stage are the vectors of probability that each frame belongs to each class. It uses this set of probability vectors to learn a track-statistical-model. The dimension of the feature space used in the second stage is therefore equal to the number of classes. In the present case it is a 105-dimensional space which has to be trained with only 20 segments of 30s which is very few. We believe this is the cause for the very low recognition rate obtained (1.11%).

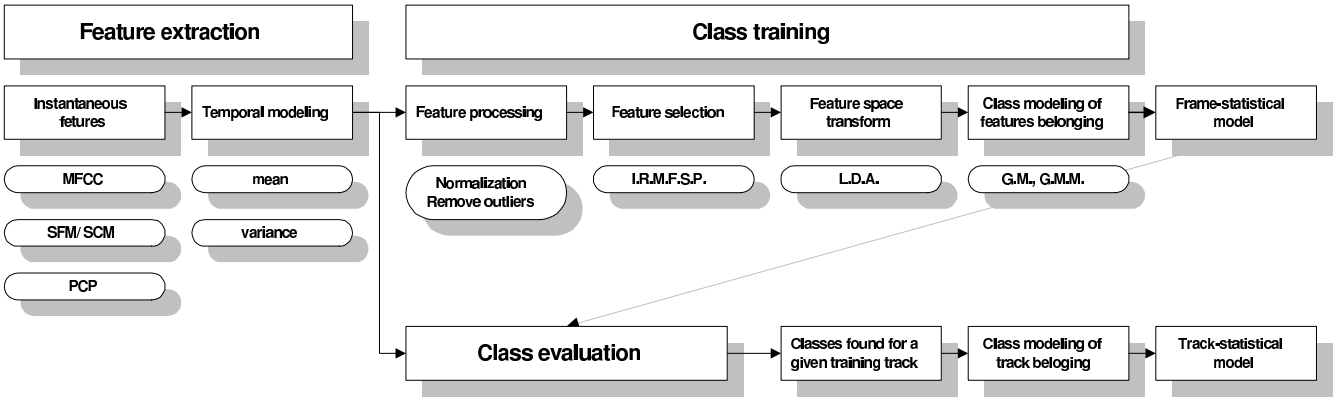

Figure 1. Flowchart of the two stages training and classification system

5.4 Audio Tag Classification

Audio Tag Classification Results										
Measure	BBE1	BBE2	BBE3	ME1	ME2	ME3	GP1	GP2	LB	TB
Average Tag Positive Example Accuracy	0.05	1.00	0.85	0.67	0.68	0.66	**0.04**	**0.03**	0.28	0.91
Average Tag Negative Example Accuracy	0.99	0.00	0.37	0.71	0.73	0.73	**0.98**	**0.98**	0.94	0.09
Average Tag F-Measure	0.06	0.15	0.19	0.24	0.26	0.26	**0.03**	**0.02**	0.28	0.15
Average Tag Accuracy	0.91	0.09	0.43	0.71	0.73	0.72	**0.90**	**0.89**	0.90	0.17
Average AUC-ROC Clip	0.82	0.49	0.81	0.77	0.79	0.78	**n/a**	**n/a**	0.84	0.69
Average AUC-ROC Tag	0.66	0.50	0.74	0.75	0.77	0.76	**n/a**	**n/a**	0.77	0.50

Table 3. Audio Tag Classification Results

The system submitted by Ircam ranked last with an F-measure around 3% (see Table 3).

One of the main characteristics of the Audio Tag Classification task is to use a set of binary classifiers, each trained using a set of Positive/ Negative examples of the tag. For a specific binary classifier, this set of examples can be very unbalanced (for some cases up to 1% Positive/ 99% Negative examples). This is problematic since the feature selection algorithm used in ircamclassification (IRMFSP), as well as the Linear Discriminant Analysis (LDA) are sensitive to the distribution of the training set. Because of this unbalancing, ircamclassification has mainly predicted Negative tags (Average Tag Positive Example Accuracy is 4% while the Average Negative Example Accuracy is 98%). We believe this unbalancing is the cause for the very low results obtained (3%).

6 ACKNOWLEDGEMENTS

This work was partly realized as part of the Quaero Programme [1] , funded by OSEO, French State agency for innovation.

7 REFERENCES

[1] T. Fujishima. Realtime chord recognition of musical sound: a system using common lisp music. In *Proc. of ICMC*, pages 464–467, Bejing, China, 1999.

[2] O. Izmirli. Using a spectral flatness based feature for audio segmentation and retrieval. In *Proc. of ISMIR*, Pymouth, Massachusetts, USA, 2000.

[3] G. Peeters. Automatic classification of large musical instrument databases using hierarchical classifiers with inertia ratio maximization. In *Proc. of AES 115th Convention*, New York, USA, 2003. Peeters03b.

[4] G. Peeters. A large set of audio features for sound description (similarity and classification) in the cuidado project. Cuidado i.s.t. report, IRCAM, 2004.

[5] G. Peeters. A generic system for audio indexing: application to speech/ music segmentation and music genre. In *Proc. of DAFX*, Bordeaux, France, 2007.

[6] G. Wakefield. Mathematical representation of joint time-chroma distributions. In *Proc. of SPIE conference on Advanced Signal Processing Algorithms, Architecture and Implementations*, pages 637–645, Denver, Colorado, USA, 1999.

[1] http://www.quaero.org