LABROSA'S AUDIO CLASSIFICATION SUBMISSIONS

Michael I. Mandel
LabROSA, Dept. Elec. Eng.
Columbia University, NY, NY
mim@ee.columbia.edu

Daniel P. W. Ellis
LabROSA, Dept. Elec. Eng.
Columbia University, NY, NY
dpwe@ee.columbia.edu

ABSTRACT

We have submitted a system to MIREX 2008's audio music classification tasks. It employs the spectral features described in [2] in addition to novel stereo-based features. For the n-way audio classification tasks (artist, classical composer, genre, latin genre, and mood identification) it uses a DAG-SVM to perform classification. For the tag classification task, it uses a simple binary SVM with Platt scaling so the SVM outputs for the different tags can be compared. Our system came in first in the audio artist and classical composer identification tasks and tied for first in many of the tag classification and retrieval tasks. A bug in the n-way classification submission prevented the comparison of features, but in the tag classification, the stereo features were shown not to be helpful.

1 SYSTEM DESIGN

This submission is very similar to last year's audio classification submissions. One difference is the inclusion of a new set of novel features characterizing the stereo content of each song. There are three versions of the submission, each using a different subset of the features, but the general structure is the same. The features are described most recently in [2], but the spectral features were originally described in [1]. The system has five main parts, the spectral, temporal, and stereo features, the similarity function, and the classifier. The system submitted to the tag classification task differs from that submitted to the other classification tasks (audio genre, mood, artist, and composer identification) in the classifier.

1.1 Features

Spectral features are the mean and covariance of a clip's MFCCs. The log of the on-diagonal variance terms is then taken to make their distribution more Gaussian. All of these features are then unwrapped to form the spectral feature vector. Technically, the spectral features are calculated on the sum of the MFCCs for the left and right channels. The stereo features are computed in exactly the same way, but on the difference of the MFCCs for the left and right channels. Because of the linearity of the DCT, the sum of the MFCCs corresponds to the sum of the log-magnitude spectra, or the product of the linear-magnitude spectra. The difference corresponds to the ratio of the linear-magnitude spectra, i.e. the interaural level difference.

The temporal features are calculated on the magnitude of the Mel spectrogram, including frequencies from 50 Hz to 10,000 Hz, using a window size of 25 ms and a hop size of 10 ms. The mel bands are added together in four large bands at low, low-mid, high-mid, and high frequencies giving the total magnitude in each band over time. The bands are windowed and their Fourier transforms are taken, from which the magnitude of the 0-10 Hz modulation frequencies are kept. The log of these magnitudes is then taken, followed by the DCT and the bottom 50 coefficients of this *envelope cepstrum* are kept. This last step boosts similarity between songs of similar rhythmic pattern, but slightly different tempo. The four bands' vectors are then stacked to form the final features.

Temporal features are calculated on 10-second clips and then averaged over overlapping windows of a sound file. For the tag classification task, the clips are 10 seconds long and this averaging is unnecessary. For the other classification tasks, the clips are 30 seconds long and so five 10-second clips that overlap 50% with their neighbors are selected and their features averaged. In this case the spectral features are also averaged, but because of their bag-of-frames assumption, this is equivalent to computing the features over the longer clip directly.

Each feature dimension is normalized to be zero mean and unit variance and then each subset of the feature vector (spectral, temporal, stereo) for each clip is normalized to be unit norm. The feature subsets are then weighted according to a set of weights that we have found empirically to work well, and the distance between all pairs of points are computed. See Table 1 for the weights used in the three versions of the algorithm. Note that the first version only uses spectral features, the second only uses spectral and temporal, and the third uses spectral, temporal, and stereo features in computing distances. A bug was found in the final version of the n-way classification submission that normalized and selected features incorrectly. In that case, the first version of the submission was least affect.

Table 1. Relative weights on the three types of features used in computing distances in the three versions of the submission

Version	Spectral	Temporal	Stereo
1	1	0	0
2	2	1	0
3	2	1	1

1.2 Similarity and classification

A gram matrix of similarities between all pairs of songs is computed from their distances. The similarity between songs i and j is calculated from their distance, d_{ij} as

$$s_{ij} = e^{-\gamma d_{ij}}, \quad (1)$$

where γ is a parameter set through cross-validation. We have found that a value of $\gamma = 1$ works well in practice.

This gram matrix is then used as the kernel for support vector machine (SVM) learning. In the multiclass classification tasks, binary SVMs are combined into a multiclass classifier using the DAG-SVM setup [4]. For tag classification, the binary SVMs are trained and used as-is. For retrieval, Platt scaling [3] is used to convert SVM outputs into estimated probabilities using a monotonic transformation. After this scaling, clips can be ranked by their probability for each tag and tags can be ranked by their probability for each clip.

2 RESULTS

The n-way classification system performed most well in the artist and classical composer identification tasks, beating out all other algorithms. In the artist identification task, it appears to have been statistically significantly better than all other algorithms except for algorithm GT2, George Tzanetakis' stereo submission. In classical composer identification, it was statistically significantly better than all algorithms except for GT2 and GP1, Geoffrey Peeter's submission. In the case of genre identification, our system performed slightly worse than GT3, George Tzanetakis' multicore submission, but many submissions were statistically similar. In the latin genre identification task, we did not perform as well as many of the other systems. Finally, in the mood identification task, GP1 performed quite well and most other submissions were indistinguishable. The bug in our submission prevented us from being able to draw meaningful conclusions about the performance of the various features.

In the tag classification task, there were many different metrics and it is difficult to say if one submission performed better overall. Our submissions performed well at the task of ranking clips by relevance to a particular tag, as measured by area under the ROC curve and it performed well at binary classification, as measured by F-measure. Our system was best by the F-measure metric, with a few statistically similar, including LB (Barrington, Turnbull, and Lankriet), but many statistically significantly worse. On the clip-retrieval task, our submission was tied for best with LB with most of the other algorithms performing significantly worse. While still performing well on the task of ranking tags for each clip, other systems performed better than ours.

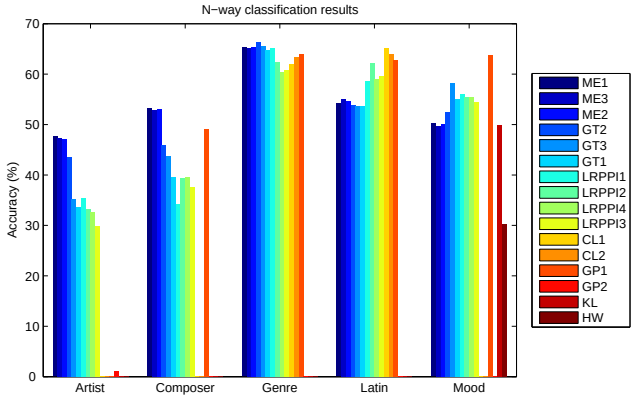
Table 2. Results of n-way classification tasks

System	Artist	Composer	Genre	Latin	Mood
ME1	47.65	53.25	65.41	54.15	50.33
ME3	47.25	52.89	65.20	54.99	49.67
ME2	47.16	53.10	65.30	54.70	50.00
GT2	43.47	45.82	66.41	53.79	52.50
GT3	35.27	43.81	65.62	53.67	58.20
GT1	33.66	39.47	64.71	53.65	55.00
LRPPI1	35.42	34.13	65.06	58.64	56.00
LRPPI2	33.20	39.43	62.26	62.23	55.50
LRPPI4	32.52	39.54	60.46	59.00	55.50
LRPPI3	29.87	37.48	60.84	59.55	54.50
CL1	—	—	62.04	65.17	—
CL2	—	—	63.39	64.04	—
GP1	—	48.99	63.90	62.72	63.67
GP2	01.11	—	—	—	—
KL	—	—	—	—	49.83
HW	—	—	—	—	30.33

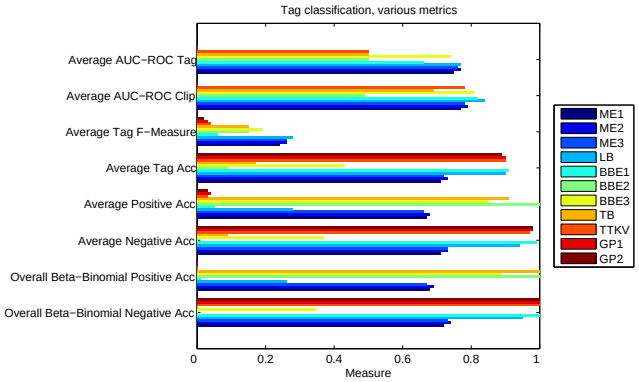
Our submission also performed consistently in terms of classification accuracy on positive and negative examples, although it is difficult to compare the accuracy results of the different systems because there are two numbers involved. These accuracies are useful, however, in measuring the operating characteristics of the various algorithms, i.e. the proportion of true positives vs false negatives and true negatives vs false positives. For example, submissions BBE1, GP1, GP2, and TTKV all had very low false positive rates, but also very low true positive rates. On the other hand BBE2 and TB had very high true positive rates, but also very high false positive rates. In the middle were our submissions, BBE3, and LB. BBE3 had a high true positive rate, but also a high false positive rate, while LB had a low true positive rate, but also a low false positive rate.

The overall binary accuracy is less informative as a metric because each tag has a different prior probability of occurrence. Differences in the overall accuracy are thus due to both classifier performance and the bias of the individual tag. In general, a system that always classified clips as negative examples of a tag would perform quite well by this metric, although it would not be useful at all in practice.

Figure 1. Results of classification tasks. (a) N-way task participants: ME: Mandel, Ellis, GT: Tzanetakis, LRPPI: Lidy, Rauber, Pertusa, Peonce de Leon, Iñesta, CL: Cao, Li, GP: Peeters, KL: Lee, HW: Wang. (b) Tag classification participants: ME: Mandel, Ellis, LB: Barrington, Turnbull, Lanckriet, BBE: Bertin-Mahieux, Bengio, Eck, TB: Bertin-Mahieux, TTKV: Trohidis, Tsoumakas, Kalliris, Vlahavas, GP: Peeters.



(a) N-way tasks



(b) Tag metrics

Of the three different versions of the features we submitted, version 2, using spectral and temporal features, but no stereo features, seemed to perform the best in general. Version 3, using stereo features in addition, did perform better than version 1, using only spectral features.

3 REFERENCES

[1] M. Mandel and D. Ellis. Song-level features and support vector machines for music classificatio n. In Joshua Reiss and Geraint Wiggins, editors, *Proc. ISMIR*, pages 594–599, 2005.

[2] M. Mandel and D. Ellis. Multiple-instance learning for music information retrieval. In *Proc. ISMIR*, September 2008. To appear.

[3] J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74, 1999.

[4] J. Platt, N. Cristianini, and J. Shawe-Taylor. Large margin DAGs for multiclass classification. In S.A. Solla, T.K. Leen, and K.-R. Mueller, editors, *Advances in Neural Information Processing Systems 12*, pages 547–553, 2000.