# LABROSA'S AUDIO MUSIC SIMILARITY AND CLASSIFICATION SUBMISSIONS

**Michael I Mandel**
Columbia University
LabROSA, Dept. Electrical Engineering
mim@ee.columbia.edu

**Daniel P W Ellis**
Columbia University
LabROSA, Dept. Electrical Engineering
dpwe@ee.columbia.edu

## ABSTRACT

We have submitted a system to MIREX 2007's audio music similarity and classification tasks. It employs spectral features based on [1] and temporal features similar to those described in [3]. For the similarity task, it calculates the distance between songs as the Euclidean distance between their feature vectors. For the audio classification tasks (artist, classical composer, genre, and mood identification) it uses a DAG-SVM [2] to perform $n$-way classification. Our system performed especially well in the audio artist and classical composer identification tasks using only the spectral features, but also performed well on the other tasks using the combined spectral and temporal features.

## 1 SYSTEM DESIGN

The system has four main parts, the spectral and temporal features, the similarity function, and the classifier.

### 1.1 Features

See Figure 1 for flowchart illustrating the feature extraction process. All of our feature extraction is performed on 10-second clips of music. To analyze MIREX's 30-second excerpts, we extract 5 overlapping 10-second clips and average the features together. For the spectral features, this averaging is equivalent to computing a single feature vector for the entire 30-second excerpt, but this is not the case for the temporal features.

The spectral features are the same as those described in [1], the mean and covariance of a clip's MFCCs. Since the on-diagonal variance terms are strictly positive, their log is taken to make their distribution more Gaussian. The covariance is unwrapped with the svec operation and stacked with the mean to form the spectral feature vector. These features are generally good at capturing timbral aspects of the music such as instrumentation and production.

The temporal features are similar to those described in [3]. They are calculated on the magnitude of the Mel spectrogram, including frequencies from 50 Hz to 10,000 KHz, 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 DCT of these magnitudes is then taken 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.

Each feature dimension is normalized over all of the clips to be zero mean and unit variance. In the classification tasks, each feature vector is further normalized to be unit norm. This renormalization invalidates the feature-wise normalization, but avoids problems with feature vectors with small norms being close to too many other vectors in the high dimensional feature space.

### 1.2 Classification and similarity

To compute the distance between two songs, the distances of their temporal and spectral features are computed separately and then combined in a weighted sum. In the primary classification and similarity submissions, labeled "ME" in Figure 2, twice as much weight was placed on the spectral as the temporal features. In the submission labeled "$\text{ME}_{\text{spec}}$", only the spectral features are used by giving the temporal features zero weight. The similarity between songs $i$ and $j$ is calculated from their distance, $d_{ij}$ as

$$s_{ij} = e^{-\gamma d_{ij}}, \tag{1}$$

where $\gamma$ is a parameter set through cross-validation.

This similarity matrix is positive definite and so can be used to define a kernel between songs for the support vector machines (SVMs) used in the classification system. Since SVMs are binary classifiers, they must be combined to perform $n$-way classification as in the classification tasks. Setting up a directed acyclic graph (DAG) of binary SVMs in the proper configuration allows training and evaluation more efficiently than other $n$-way schemes [2].

## 2 RESULTS

The classification system performed most well in the artist and classical composer identification tasks. Only IMIRSEL's entry performed better, using similar features and a similar

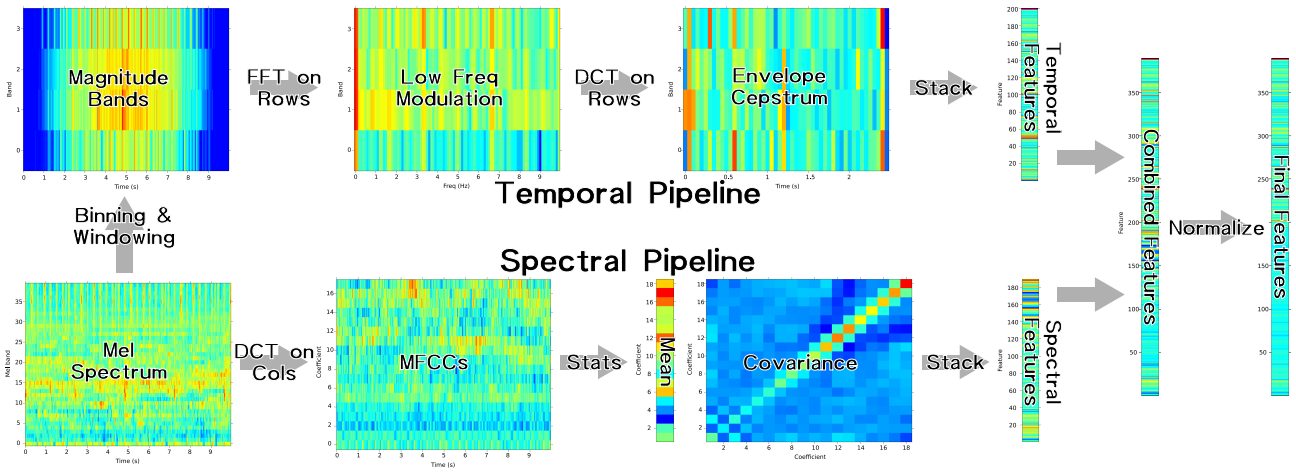
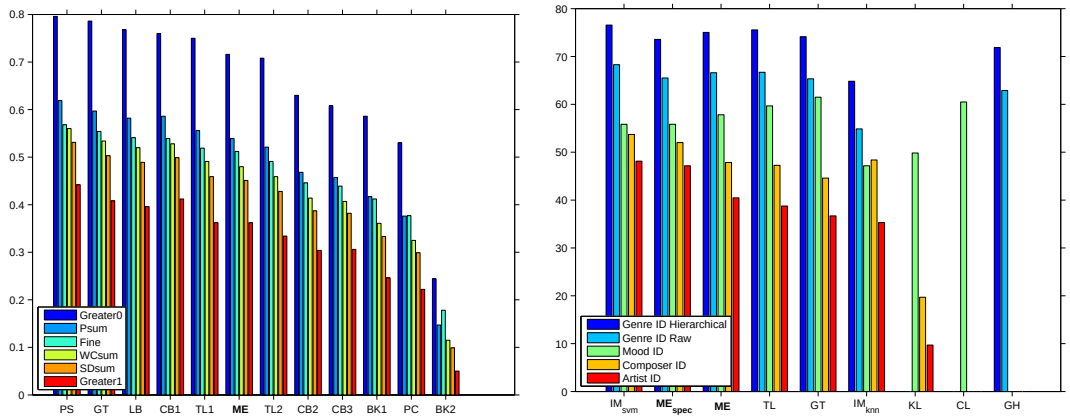



**Figure 1**. Flowchart of the feature extraction process. The example song is Delerium's "Consentual Worlds" from their album *Semantic Spaces*.

**Figure 2**. Results of MIREX similarity and classification evaluations




(a) Audio Music Similarity

BK = Bosteels, Kerre; CB = Christoph Bastuck; GT = George Tzanetakis; LB = Barrington, Turnbull, Torres, Lanckriet; ME = Mandel, Ellis; PC = Paradzinets, Chen; PS = Pohle, Schnitzer; TL = Lidy, Rauber, Pertusa, Iñesta

(b) Audio Classification

CL = Laurier, Herrera; GH = Guaus, Herrera; GT = George Tzanetakis; IM = IMIRSEL M2K; KL = Kyogu Lee; ME = Mandel, Ellis; TL = Lidy, Rauber, Pertusa, Manuel Iñesta

classifier. In these two tasks the temporal features actually hurt performance, although they helped in the genre and mood classification tasks. Results on the genre and mood tasks are similar to the similarity tasks, but rather dissimilar from those of the artist and composer identification tasks.

The similarity system finished in the middle of the pack. The difference between its performance and the top systems' performance was just barely statistically significant, according to IMIRSEL's Friedman test. As can be seen in Figure 2(a), the rankings produced by all of the metrics were quite consistent.

**Run-time** Feature extraction dominates the runtime of our system. Once features are calculated, training, classifying, and similarity determination are very fast. On a 3GHz Xeon, feature extraction ran about 10 times faster than real time, and the other operations ran in less than a second each. As the size of the database grows, however, these other operations start to dominate the runtime. For $n$ songs, similarity calculations are $O(n^2)$ in time and for $n$ training songs, SVM training is $O(n^3)$ in time.

## 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] John C. Platt, Nello Cristianini, and John 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.

[3] Andreas Rauber, Elias Pampalk, and Dieter Merkl. Using psycho-acoustic models and self-organizing maps to create a hie rarchical structuring of music by sound similarity. In Michael Fingerhut, editor, *Proc. ISMIR*, pages 71–80, 2002.