MIREX 2007: AUDIO COVER SONG DETECTION USING CHROMA FEATURES AND A HIDDEN MARKOV MODEL

Youngmoo E. Kim
Drexel University
Electrical & Computer Eng.

Daniel Perelstein
Swarthmore College
Engineering Department

ABSTRACT

Our submission to the MIREX 2007 Audio Cover Song Detection task uses Hidden Markov Models trained from pitch chroma features estimated from the sound data. We hypothesize that the state sequences generated from the HMMs from cover songs will be similar to each other, at least in terms of the relative frequency of the states, and we use the state histograms to assess the relative distance between one song and another.

1 INTRODUCTION

Different performances of a song will likely have very different acoustic properties from each other, including transposition of key, changes in instrumentation, variations in tempi, etc. Several factors identify the underlying composition, including the general pattern of harmonies, as well as the song structure and lyrics. Our system focuses on the harmonic patterns of the piece, first using an acoustic representation based on relative changes in pitch chroma as the primary feature. The patterns of variation between harmonies are then captured using a Hidden Markov Model, which is designed to model the time-variation of different harmonic states.

2 ACOUSTIC FEATURE EXTRACTION

Our system first calculates a relative pitch chroma representation from the audio waveform. A well-established method for estimating the pitch chroma components within a short time-interval of audio is the chromagram [1]. The chromagram is essentially a circular version of the spectrogram, where the frequencies of chroma in different octaves are grouped together and summed to provide the summary of energy at each of the 12 pitch classes. Our chroma features are calculated at a constant frame rate of $\sim$0.1 sec, and then smoothed over a short-time window ($\sim$0.65 sec) that gives the greatest weight to the most recent frames.

Because the different performances are unlikely to all be in the same musical key, it is necessary to use a relative representation that attempts to capture the change in chroma spectra from one moment in time to the next. It is, however, difficult to accurately (and automatically) detect when a harmonic change occurs. We attempt to model the relative changes by performing a cross-correlation of each chroma frame with each of the preceding 20 frames (a window of $\sim$1.8 sec) for each possible chroma interval (11 possible shifts). The cross-correlations are then averaged across the time window to form the overall feature vector, which we call *cross-correlated chromagrams* (CCC). Since the correlation time-window is fairly large, a change in component chroma be reflected over some time. It is hoped that the CCC frames are representative of the relative changes in chroma (and thus harmony) over time.

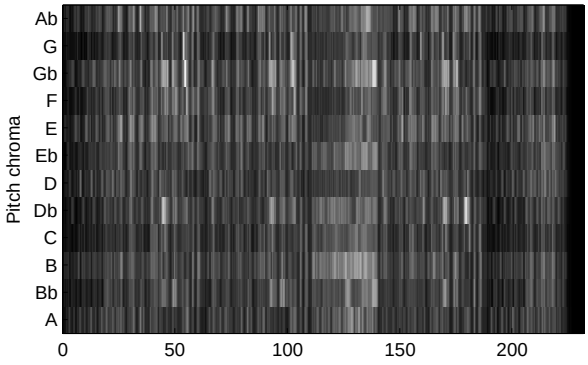

Figure 1. Smoothed chromagram of an example song.

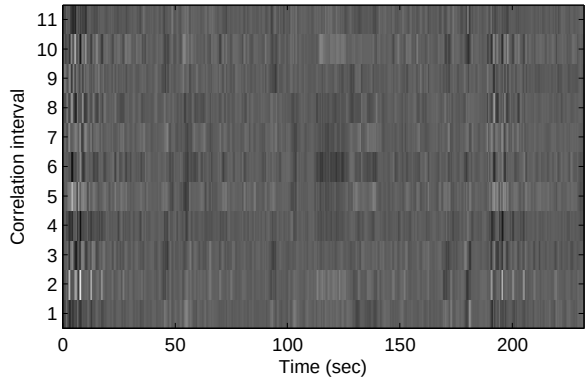

Figure 2. Cross-correlated chromagram.

3 MODEL TRAINING

The CCC features are used to train a Hidden Markov Model (HMM) [2] whose states should reflect the relative harmonic changes over time (including when there is no change). This initial "harmonic" model consists, somewhat arbitrarily, of 35 states. Each state's observations (the 11-dimensional CCC features) are modeled as a single multivariate Gaussian distribution.

Our model parameters were trained using the standard forward-backward (Expectation-Maximization) algorithm using features collected from a randomly chosen subset of audio files from a large database of popular music. Once the parameters have been trained, features from new songs can be evaluated using the trained HMM into a most likely sequence of states (MLSS). Since it is hoped that songs with similar harmonies will have similar state sequences, the histograms of MLSS vectors from each song are calculated and used as the basis of comparison between songs.

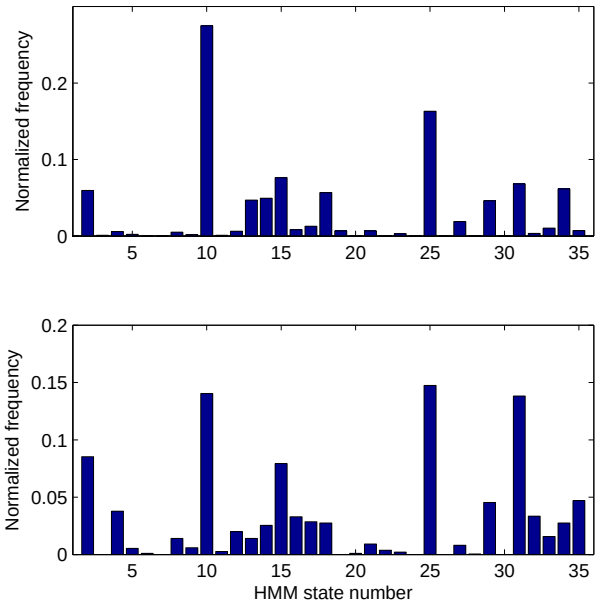

Figure 3. Comparison of histograms of maximum likelihood state sequence from two covers of the same song.

4 DISTANCE EVALUATION

The histogram from each song's MLSS is normalized to be a unit norm vector and is compared to that of other songs using a simple dot product (the greater the similarity between the two histograms vectors, the greater the result).

5 REFERENCES

[1] M. A. Bartsch and G. H. Wakefield, "To catch a chorus: Using chroma-based representations for audio thumbnailing," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, October 2001.

[2] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, February 1989.