CLASSIFYING MUSIC AUDIO WITH TIMBRAL AND CHROMA FEATURES

Daniel P. W. Ellis
LabROSA, Dept. Elec. Eng.
Columbia University

ABSTRACT

Music audio classification has most often been addressed by modeling the statistics of broad spectral features, which, by design, exclude pitch information and reflect mainly instrumentation. We investigate using instead beat-synchronous chroma features, designed to reflect melodic and harmonic content and be invariant to instrumentation. Chroma features are less informative for classes such as artist, but contain information that is almost entirely independent of the spectral features, and hence the two can be profitably combined: Using a simple Gaussian classifier on a 20-way pop music artist identification task, we achieve 54% accuracy with MFCCs, 30% with chroma vectors, and 57% by combining the two. All the data and Matlab code to obtain these results are available. [1]

1 INTRODUCTION

Classifying music audio (for instance by genre or artist) has been most successful when using coarse spectral features (e.g. Mel-Frequency Cepstral Coefficients or MFCCs [3]), which has been dubbed "timbral similarity" [1]. Such features reflect mainly the instruments and arrangements of the music rather than the melodic or harmonic content. Although a wide range of more musically-relevant features has been proposed, they have rarely afforded much improvement on the overall system performance.

This paper describes the results of using beat-synchronous chroma features in place of frame-level MFCCs in statistical classification of artist identification. This feature representation was developed for matching "cover songs" (alternative performances of the same musical piece), and thus reflects harmonic and melodic content with the minimum of influence from instrumentation and tempo [2]. Chroma features consist of a twelve-element vector with each dimension representing the intensity associated with a particular semitone, regardless of octave; our implementation uses instantaneous frequency to improve frequency resolution and rejection of non-tonal energy. By storing just one chroma vector for each beat-length segment of the original audio (as determined by a beat tracker), the representation is both compact and somewhat tempo invariant – yet still sufficient, since chroma content rarely changes within a single beat, provided the beats are chosen near the bottom of the metrical hierarchy.

Using chroma features directly in place of MFCCs means that we are looking only at the global distribution of chroma use within each piece, and assuming this is somewhat consistent within classes. A chroma vector reflects all current notes and is thus roughly correlated with a particular chord (such as Bmin7). Thus, the kind of regularity that such a model might capture would be if a given artist was particularly fond of a certain subset of chords. The model in this form would not learn characteristic chord sequences, and would also be defeated by simple transposition.

To overcome these problems we tried (a) modeling sequences of chroma vectors for several beats, and (b) using a key-normalization front-end that attempts to align every piece to a common chroma transposition.

Although we believe that beat-chroma distribution models can capture something specific about individual composition style, we do not expect them to perform as well as timbral features, since artists are likely to vary their melodic-harmonic palette more readily than their instrumentation. However, we expect the information from these two sources to be complementary, since harmonic "signatures", to the extent they exist, have no reason to be associated with instrumentation. For this reason, we experiment with fusing the two results in final classification.

2 CLASSIFIER

We adopt an artist identification task to illustrate the utility of beat-chroma features. We opt for the simple approach of fitting the distributions of random subsets of pooled frame-level features for each artist with either a single full-covariance Gaussian, or a mixture of diagonal-covariance Gaussians. Classification of an unknown track is achieved by finding the model that gives the best total likelihood for a random subset of the features from that track (treated as independent). Although previous work has shown that the SVM classifier applied to track-level features is a more powerful approach to this problem [3], our interest here is in comparing the usefulness of the different features. The simple and quick Gaussian models are adequate for this purpose.

[1] `http://labrosa.ee.columbia.edu/projects/timbrechroma/`

3 KEY NORMALIZATION

The goal of key normalization was to find a per-track rotation of the circular, 12-bin chroma representation that made all the data within each training class as similar as possible. To do this we first fit a single, full-covariance Gaussian to the chroma representation of first track in the class. Then the likelihood of each subsequent track was evaluated using this model under each of the 12 possible chroma rotations; the most likely rotation was retained. After one pass through the entire set, a new Gaussian was fit to all the tracks after applying the best rotations from the first pass. The search for the best rotations was repeated based on this new model, and the process iterated until no changes in rotations occurred. The final, global model was also used on test tracks to choose the best rotation for them prior to classification.

4 DATA

We collected and used a set of 1412 tracks, composed of six albums from each of 20 artists. It is based on the 18 artist set used on [3] (drawn from "uspop2002") with some additions and enhancements. We used 6-fold testing, with five albums from each artist used for training and one for testing in each fold. For this test set, statistical significance at 5% requires a difference of just over 3% in classification accuracy (one-tailed binomial). Guessing the most common class gives a baseline accuracy of 6%.

5 EXPERIMENTS

For both MFCC and beat-chroma features, we experimented with varying the number of frames used to train and test the models, and the number of Gaussians used to model the distributions. We used 20 MFCC coefficients including the zero'th, based on a 20-bin Mel spectrum extending to 8 kHz. For the beat-chroma features, we varied the number of temporally-adjacent frames modeled from 1 to 4, and with using key normalization.

The results are summarized in table 1. We notice that MFCCs are intrinsically more useful than chroma features (as expected, since the instrumentation captured by MFCCs is well correlated with artist), that Gaussian mixture models are preferable for the chroma features (which are likely to be multimodal) but not for MFCCs, that key normalization gives a small but significant improvement, and that concatenating multiple beats into a single feature vector gives small but consistent advantages for the chroma features. [2] Fusing the best MFCC and Chroma systems (shown in bold in the table) based on a weighted sum of separate model likelihoods tuned on a small tuning subset, we see a statistically significant improvement that results from including chroma information.

[2] This conclusion is reversed from the original and published version of this paper after a bug was found in the implementation of concatenating multiple beats.

Feature	Model	T win	Acc	Exec time
MFCC20	FullCov	1	**56%**	127 s
MFCC20	64 GMM	1	56%	563 s
Chroma	FullCov	1	14%	21 s
Chroma	FullCov	4	20%	57 s
Chroma	64GMM	1	25%	337 s
Chroma	64GMM	4	29%	1060 s
ChromaKN	FullCov	1	23%	70 s
ChromaKN	FullCov	4	28%	197 s
ChromaKN	64GMM	1	32%	516 s
ChromaKN	64GMM	4	**33%**	1238 s
MFCC + Chroma fusion			**59%**	

Table 1. Artist identification accuracy. "T win" is the size of temporal context window. "FullCov" designates single, full-covariance Gaussian models. "ChromaKN" refers to per-track key-normalized chroma data. Execution times are per fold on a 1.8 GHz Xeon E5310 CPU.

6 CONCLUSIONS

We have shown that the simple approach of modeling the distribution of per-beat chroma vectors very much as cepstral vectors have been modeled in the past is a viable approach for artist identification. Rotating the chroma vectors in order to transpose all pieces to a common tonal framework is necessary to realize the full benefits, and concatenating several frames appears to offer some advantage.

Our future plans are to pursue the idea of modeling small fragments of beat-chroma representation to identify the most distinctive and discriminative fragments characteristic of each composer/artist.

7 ACKNOWLEDGEMENTS

This work was supported by the Columbia Academic Quality Fund, and by the NSF under Grant No. IIS-0238301. Any opinions, findings, etc. are those of the authors and do not necessarily reflect the views of the NSF.

8 REFERENCES

[1] Jean-Julien Aucouturier and Francois Pachet. Music similarity measures: What's the use? In *Proc. 3rd International Symposium on Music Information Retrieval ISMIR*, Paris, 2002.

[2] D. P. W. Ellis and G. Poliner. Identifying cover songs with chroma features and dynamic programming beat tracking. In *Proc. ICASSP*, pages IV–1429–1432, Hawai'i, 2007.

[3] M. I. Mandel and D. P. W. Ellis. Song-level features and support vector machines for music classification. In *Proc. International Conference on Music Information Retrieval ISMIR*, pages 594–599, London, Sep 2005.