CROSS-CORRELATION OF BEAT-SYNCHRONOUS REPRESENTATIONS FOR MUSIC SIMILARITY

Daniel P. W. Ellis, Courtenay V. Cotton, and Michael I. Mandel

LabROSA, Dept. of Electrical Engineering
Columbia University, New York NY 10027 USA
{dpwe,cvcotton,mim}@ee.columbia.edu

ABSTRACT

Systems to predict human judgments of music similarity directly from the audio have generally been based on the global statistics of spectral feature vectors i.e. collapsing any large-scale temporal structure in the data. Based on our work in identifying alternative ("cover") versions of pieces, we investigate using direct correlation of beat-synchronous representations of music audio to find segments that are similar not only in feature statistics, but in the relative positioning of those features in tempo-normalized time. Given a large enough search database, good matches by this metric should have very high perceived similarity to query items. We evaluate our system through a listening test in which subjects rated system-generated matches as similar or not similar, and compared results to a more conventional timbral and rhythmic similarity baseline, and to random selections.

Index Terms— Music, Database searching, Acoustic signal analysis, Dynamic programming, Correlation

1. INTRODUCTION

A system that could accurately predict judgments of musical similarity based only on the audio signals would have a number of interesting applications in helping listeners to find music that might interest them without requiring intermediaries such as record companies or radio stations. This task has attracted significant attention in recent years, with the most successful systems relying mainly on statistics of spectral feature vectors such as MFCCs [1, 2]. There has been a series of formal evaluations of these systems, although in many cases a proxy task such as genre classification or artist identification is used since establishing the 'right' answer for subjective judgments of similarity is a significant challenge [3]. However, starting in 2006, these evaluations have also included subjective evaluation of the returns from similarity systems [4].

Subjective similarity is likely to be very difficult to predict because it may be based on any number of aspects of the music. While current systems are largely sensitive to the instrumentation and production style (which is well correlated with genre), listeners may judge songs as similar because of melodic or harmonic similarity, common themes in the lyrics, or even indirect factors such as some link between the artists responsible for each piece that is most likely not discernible from the audio itself.

At a seeming tangent to this work, we have been looking at the problem of identifying "cover songs" – alternative versions of musical works, usually by different performers, that modify the interpretation in terms of instrumentation, tempo, style, or even harmonization. Our cover song identification system is very successful at identifying cover versions of popular music tracks even when the new interpretation is radically different [5]. The system operates by tracking the beat of the music audio being analyzed, storing for each beat a 12-dimensional 'chroma' representation that captures the relative intensity of the pitches corresponding to the 12 distinct semitones in the western octave, then cross-correlating these beat-chroma representations between pieces to find time alignments (and possibly musical transpositions) that lead to high similarity.

Sometimes, two pieces of music can have a striking similarity because of the coincidental use of a particular melodic or instrumentational motif. We were curious whether a system built to find such shared patterns would be able to find near matches that would also strike a listener as similar. In the context of the subjective evaluations mentioned above, it is possible that a near-match to a particular melodic, harmonic, or instrumental sequence – the kind of thing our cross-correlation approach might find – would give a much stronger sense of similarity than the overall statistical resemblance being found by the current similarity systems. Hence, in this work, we adapted our cover song system to find the most similar clips from a database which is presumed not to include any cover versions of the query clips, but which may contain other fragments that happen to be relatively close to what a cover version might be like.

This work was supported by the NSF (grant IIS-0713334) and the Columbia Academic Quality Fund.

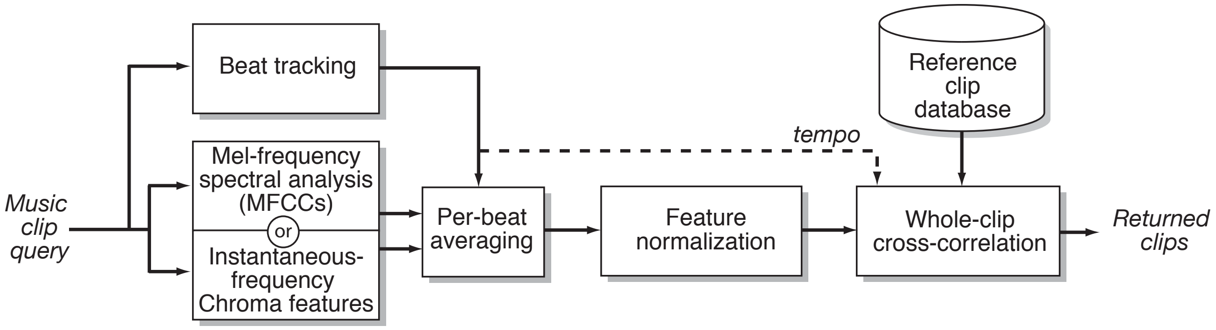

Fig. 1. Block diagram of the cross-correlation based music similarity system

2. CROSS-CORRELATION SIMILARITY

Figure 1 shows the block diagram of our similarity system, which is largely the same as our cover song system [5]. First, a dominant rhythmic pulse is found by auto-correlation, and then an optimal sequence of beat times is found by dynamic programming [6]. Base features are calculated on a uniformly-sampled time base (e.g. every 10 ms), then averaged within each beat segment to construct the beat-synchronous feature representation. In our original cover song system, the features were the chroma vectors, calculated by identifying tonal components in the spectrum below 2 kHz, and mapping their precise frequencies (from the instantaneous frequency) to one of the twelve chroma bins. In this work we also experimented with using the conventional Mel-Frequency Cepstral Coefficients (MFCCs) in exactly the same role, reasoning that a particular sequence of spectral variation may be as important, or even more important, in leading to a subjective impression of similarity. We have also looked at combining similarity based on each feature.

Beat-synchronous feature matrices (12 rows by the number of beats identified) are then normalized to have zero mean and unit variance within each dimension; they are also high-pass filtered along time to emphasize changes, since these seem more perceptually salient. This compact representation is pre-calculated for all items in the reference database as well as being calculated for each query clip. Matching the query clip consists of evaluating the full cross-correlation between the representation of the query and every item in the reference database. Although this involves a cross-correlation for each reference clip, these can be performed via the fast Fourier transform and can thus be partially pre-computed. (In the case of chroma features, a circular cross-correlation is additionally performed on the 'vertical' axis to find the best relative transposition). The peak value of the cross-correlation, without further normalization, is taken as a measure of similarity between query and reference items, and the reference items can then be ranked by similarity to the query. [1]

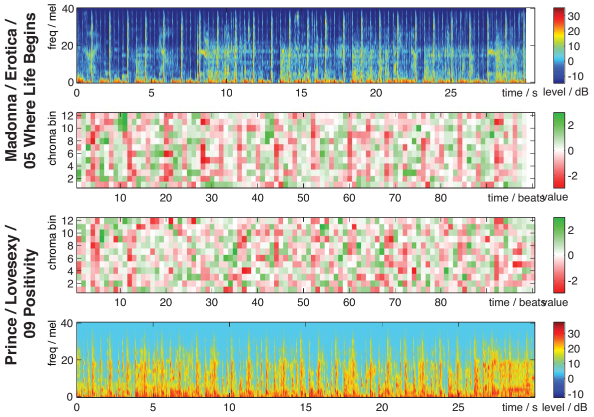

Fig. 2. Example match between the beat-synchronous MFCC representations of two tracks which are judged as similar even though they are unrelated musical pieces. The upper panes show the Mel-frequency spectrogram, and normalized beat-synchronous MFCC matrix for a Madonna excerpt. The lower panes show the aligned representation of a Prince excerpt.

Figure 2 shows an example of two short clips that matched under this approach, using MFCCs as the base features, at their best alignment. Note the visible 8-beat repetition aligned in both excerpts.

2.1. Boundary alignment

Although using FFTs to compute cross-correlation allows an efficient implementation, it still requires significant computation since the inner products between the feature arrays must be computed for every possible time skew. If, instead, the 'best' candidate time skews were identified ahead of time, computation could be dramatically reduced by comparing only at those few alignments. Indeed, the entire search for matches to a query could then be reduced to a nearest-neighbor search, which can be implemented in sub-linear time via locality-sensitive hashing as suggested in [7]. To evaluate the viability of such an approach, we implemented a simple system

[1] The code for the similarity system is available at `http://labrosa.ee.columbia.edu/projects/xcorrsim/`.

that assigns a single time anchor or boundary within each segment, then calculates the correlation only at the single skew that aligns the time anchors. We use the BIC method [8] to find the boundary time within the feature matrix that maximizes the likelihood advantage of fitting separate Gaussians to the features each side of the boundary compared to fitting the entire sequence with a single Gaussian i.e. the time point that divides the feature array into maximally dissimilar parts. While almost certain to miss some of the matching alignments, an approach of this simplicity may be the only viable option when searching in databases consisting of millions of tracks.

3. EXPERIMENTS AND RESULTS

The major challenge in developing music similarity systems is performing any kind of quantitative analysis. As noted above, the genre and artist classification tasks that have been used as proxies in the past most likely fall short of accounting for subjective similarity, particularly in the case of a system such as ours which aims to match structural detail instead of overall statistics. Thus, we conducted a small subjective listening test of our own, modeled after the MIREX music similarity evaluations [4], but adapted to collect only a single similar/not similar judgment for each returned clip (to simplify the task for the labelers), and including some random selections to allow a lower-bound comparison.

3.1. Data

Our data was drawn from the uspop2002 dataset of 8764 popular music tracks. We wanted to work with a single, broad genre (i.e. pop) to avoid confounding the relatively simple discrimination of grossly different genres with the more subtle question of similarity. We also wanted to maximize the density of our database within the area of coverage.

For each track, we took a 10 s excerpt from 60 s into the track (tracks shorter than this were not included). We chose 10 s based on our earlier experiments with clips of this length that showed this is an adequate length for listeners to get a sense of the music, yet short enough that they will probably listen to the whole clip [9]. (MIREX uses 30 s clips which are quite arduous to listen through).

3.2. Comparison systems

Our test involved rating ten possible matches for each query. Five of these were based on the system described above: we included (1) the best match from cross-correlating chroma features, (2) from cross-correlating MFCCs, (3) from a combined score constructed as the harmonic mean of the chroma and MFCC scores, (4) based on the combined score but additionally constraining the tempos (from the beat tracker) of database items to be within 5% of the query tempo, and (5) combined score evaluated only at the reference boundary of section 2.1. To these, we added three additional hits from a more conventional feature statistics system using (6) MFCC mean and covariance (as in [2]), (7) subband rhythmic features (modulation spectra, similar to [10]), and (8) a simple summation of the normalized scores under these two measures. Finally, we added two randomly-chosen clips to bring the total to ten.

Table 1. Results of the subjective similarity evaluation. Counts are the number of times the best hit returned by each algorithm was rated as similar by a human rater. Each algorithm provided one return for each of 30 queries, and was judged by 6 raters, hence the counts are out of a maximum possible of 180.

Algorithm	Similar count
(1) Xcorr, chroma	48/180 = 27%
(2) Xcorr, MFCC	48/180 = 27%
(3) Xcorr, combo	55/180 = 31%
(4) Xcorr, combo + tempo	34/180 = 19%
(5) Xcorr, combo at boundary	49/180 = 27%
(6) Baseline, MFCC	81/180 = 45%
(7) Baseline, rhythmic	49/180 = 27%
(8) Baseline, combo	**88/180 = 49%**
Random choice 1	22/180 = 12%
Random choice 2	28/180 = 16%

3.3. Collecting subjective judgments

We generated the sets of ten matches for 30 randomly-chosen query clips. We constructed a web-based rating scheme, where raters were presented all ten matches for a given query on a single screen, with the ability to play the query and any of the results in any order, and to click a box to mark any of the returns as being judged “similar” (binary judgment). Each subject was presented the queries in a random sequence, and the order of the matches was randomized on each page. Subjects were able to pause and resume labeling as often as they wished. Complete labeling of all 30 queries took around one hour total. 6 volunteers from our lab completed the labeling, giving 6 binary votes for each of the 10 returns for each of the 30 queries.

3.4. Results

Table 1 shows the results of our evaluation. The binary similarity ratings across all raters and all queries are pooled for each algorithm to give an overall ‘success rate’ out of a possible 180 points – roughly, the probability that a query returned by this algorithm will be rated as similar by a human judge. A conservative binomial significance test requires a difference of around 13 votes (7%) to consider two algorithms different.

We see that the MFCC-based comparison system performs far above any other single feature, and the addition of rhythmic features further improves it slightly (although not enough to achieve significance on this test). It is interesting to see the random choices score as well as they did; most likely, there is a complex procedure by which the raters adjust their working definition of "similarity" based on the range of qualities of matches they encounter, and it may have been stretched quite far. However, the cross-correlation systems all performed significantly better than the random baseline with the exception of (6), the tempo-filtered results; we infer that the tempo filtering reduced the potential pool of matches too far. A further analysis of which algorithm performed best (collected the most similar ratings) for each individual query had the cross-correlation systems outperform the comparison systems in 10 out of 30 queries, indicating that even though it is less successful overall, it is able to find some matches that can better those found by conventional methods. Also, of the 18 returned items considered similar to their queries by all six raters, 7 were returns from the cross-correlation approaches.

Fleiss's kappa for the inter-rater agreement was 0.36 suggesting fair consistency. Raters will differ based on their individual music taste and familiarity with particular styles.

4. DISCUSSION AND CONCLUSIONS

It is interesting to note that the problem we are trying to solve is mainly about *precision*: in the evaluation, the listener only rates the actual items returned by the system. It does not consider *recall* – if there are other items in the database that should have been returned, the listener will not know of their existence and cannot penalize the system for missing them. Secondly, there is likely a nonlinear interaction with the size of the database being searched: as the database grows in size, it becomes increasingly likely that any given query clip will have a cover-like match somewhere in the database. If even the best melodic match to a query is not all that similar, the listener will probably not notice the similarity – in a sense, there is a 'critical radius' of similarity, outside of which the cover-style matches are not judged similar. We do not yet know what kind of database size is required to achieve sufficient density to make it likely to find matching clips within this radius.

Although the results of this experiment showed that the conventional approach of matching the statistics of spectral features after collapsing across time is a more effective approach, we have shown that direct comparison of beat-level features can find some matches that are judged similar by listeners. We remain very interested in the possibility of using similarity in beat-synchronous feature sequences for finding perceptually-close matches from very large databases. In particular, the boundary-based technique was only slightly inferior to full cross-correlation (below statistical significance in our tests), so the possibility of using very fast hashing techniques to allow searches of this kind in databases several orders of magnitude larger than the 8800 clips used here gives the promise of much denser coverage and the greater likelihood of uncannily similar matches being discovered.

5. REFERENCES

[1] J.-J. Aucouturier and F. Pachet, "Music similarity measures: What's the use?," in *Proc. 3rd International Symposium on Music Information Retrieval ISMIR*, Paris, 2002.

[2] 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*, London, Sep 2005, pp. 594–599.

[3] J. Downie, K. West, A. Ehmann, and E. Vincent, "The 2005 Music Information Retrieval Evaluation eXchange (MIREX 2005): Preliminary overview," in *Proceedings of the International Conference on Music Information Retrieval*, London, 2005, pp. 320–323.

[4] A. A. Gruzd, J. S. Downie, M. C. Jones, and J. H. Lee, "Evalutron 6000: collecting music relevance judgments," in *Proc. Joint Conference on Digital Libraries (JCDL)*, Vancouver, BC, 2007, p. 507.

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

[6] D. P. W. Ellis, "Beat tracking by dynamic programming," *J. New Music Research*, 2007, Special Issue on Tempo and Beat Extraction, to appear.

[7] M. Casey and M. Slaney, "The importance of sequences in musical similarity," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toulouse, 2006, pp. V–5–8.

[8] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

[9] M. I. Mandel and D. P. W. Ellis, "A web-based game for collecting music metadata," in *Proc. International Conference on Music Information Retrieval ISMIR*, Vienna, 2007.

[10] A. Rauber, E. Pampalk, and D. Merkl, "Using psychoacoustic models and self-organizing maps to create a hierarchical structuring of music by sound similarities," in *Proc. Int. Symposium on Music Information Retrieval (ISMIR)*, Paris, 2002.