Data-Driven Music Audio Understanding
IIS-0713334
Annual Report 2008

Daniel P.W. Ellis
Department of Electrical Engineering,
Columbia University
500 W. 120th Steet, New York NY 10027, USA

dpwe@ee.columbia.edu

Sep 08, 2008

1 Activities & Findings

1.1 Research Activities

This project is concerned with extracting latent information from large collections of music audio – for instance, to answer the question of what common characteristics music audio examples share that can differentiate them from non-music, or unpleasant music. To this end, we have been developing music audio analysis techniques at several levels:

1.1.1 Music transcription and recognition

At the lowest, most literal level, we have continued our work on describing the moment-to-moment development of music audio either in the conventional terms of music notation (notes, chords etc.), or in closely related features such as chroma vectors (the 'weight' of a moment as distributed among the 12 semitones of one octave on the piano keyboard).

For note-level transcription, we have investigated developing our previous work on trained classifier-based transcription to take advantage of the large volumes of unlabeled data (previously, our system was trained only on data for which labels where obtained) [7]. In the case where a nominal score is available (i.e. a list of the canonical note events, but without performance nuance in timing, tuning, or accents), we are developing a system for high-resolution alignment of score to performance, able to measure fine timing and intonation modifications that may have interesting musicological implications, picking out single voices even in polyphonic and reverberant contexts [1].

Chroma representations, which describe audio spectra by collapsing periodic energy across octaves into 12 semitone bins spanning a single octave, present a promising intermediary between information that can be robustly extracted from audio, and a representation that can be closely related to musical features of melody and harmony. We have continued to develop our beat-synchronous chroma representation, including using it as the basis for a chord recognition system that was entered into the 2008 MIREX Audio Chord Detection evaluation [3].

1.1.2 Music classification and tagging

A higher-level approach to the automatic 'understanding' of music by computers is to look at classifying, labeling, or organizing music audio at the level of short excerpts or entire tracks – comprising many individual notes, beats, and chords – based on statistical measures of the excerpt. We have continued our work on developing automatic music classification systems based both on the statistics of broad spectral properties as captured by the MFCC features used in speech recognition, as well as for the melody/harmony-related chroma features mentioned above [2, 4].

There has been significant interest recently in the use of 'tags' to describe multimedia content – for instance, the words or brief phrases that users may attach to individual music tracks to help them organize their own collections, or otherwise as an aid to retrieval. This is a inviting area for automation, since human ground-truth data can be obtained, and the task of assigning discrete tag-labels to music audio excerpts fits nicely within the classical pattern recognition framework. We have developed a web-based game that involves listeners tagging 10 s music excerpts (the object of the game is to reproduce the tags given by other players, or to have other players match yours), and we have used the data from this game to train automatic classifiers `http://www.majorminer.org` [5].

One specific issue that arises with these data is deciding whether the tag ap-

plies to the entire excerpt, or only to particular time ranges within the excerpt. This general problem, of deciding how a label applied to a collection of items is actually distributed among the individual items, is known as "multiple-instance learning" in the machine learning community, and has attracted a number of sophisticated approaches. We cast our music labeling problem in this framework (for instance, a track tagged with "saxophone" may in fact contain only a saxophone solo at a certain point in the song, and for the rest of the song the label is not relevant), and investigated using and adapting some of the existing algorithms for improving the accuracy of automatic tagging ("autotagging") [6].

1.1.3 Music data mining

The heart of this project is the idea of using analyses and representations of large music corpora to reveal and circumscribe more abstract, general properties of the music itself. This year we have conducted a number of preliminary investigations along these lines. We started by making a beat-synchronous chroma analysis of a sizable collection of music audio (our "artist20" set of 1,413 tracks by 20 different contemporary pop music artists). We then chopped these representations into fixed-length segments (e.g. 24 beats by 12 chroma bins), either at every beat, or around a series of "segmentation points" automatically identified as points of maximum contrast in feature statistics, intended to find the beginnings of verse/chorus segments etc. This resulted in roughly a million data items, including transpositions in an attempt to remove the confound of different keys. We then attempted to identify clusters of similar data items by building a Locality-Sensitive Hash (LSH) index of all the items, then performing a search for all items within a fixed (euclidean) radius of every item in the database to identify the centroids of the most densely populated regions of this high-dimensional space. LSH allows each of these queries to be answered in constant time, meaning that the entire search took only minutes. The result is a number of beat-chroma patterns that occur with relatively small variations in multiple, otherwise unrelated pieces. Our ongoing goal is to look for ways to normalize away insignificant surface variations in the representations from different pieces (due, for instance, to amplitude variations between notes that might be realizing the same underlying chord) to be able to cluster similarities that occur below the most evident, surface level of the music.

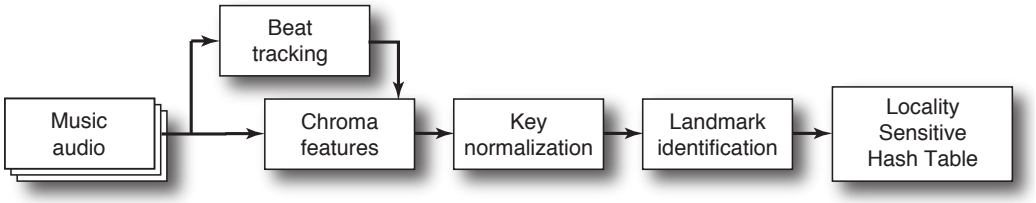

Figure 1: Block diagram of the system for clustering beat-synchronous chroma fragments extracted from music audio.

1.2 Educational activities

As part of this project, we have continued our weekly interdepartmental seminar course, where music composition students working with computers and electronics, and electrical engineering students looking at computer analysis of music audio, meet and share ideas – the Music Engineering Art Project, or MEAP. This year, we worked slowly towards the idea of a kind of music installation, the MEAPbira, which will be a computer-driven musical instrument consisting of 12 metal tines tuned to the 12 semitones of an octave (similar to the mbira musical instrument of Zimbabwe). The idea is to have a kiosk with a minijack cable coming out; a listener can plug it into their own personal stereo, play in some music, and the MEAPbira will then re-render a (hopefully) recognizable version of what it hears by plucking the metal tines with servo motors. The analysis will involve beat tracking and chroma analysis, as described above.

This project has also fed ideas and demonstrations into other ongoing courses, specifically the graduate class on Speech and Audio Processing and Recognition taught by PI Ellis. Music audio analysis now accounts for roughly one-quarter of the total course content.

PI Ellis has also supervised a number of one-semester individual student projects on this material, including several related to the graduate class mentioned above, and another on music audio segmentation with an enthusiastic undergraduate (junior), Vishal Kumar.

1.3 Findings

The findings from the research projects listed above include:

- We were able to improve the performance of our classification-based piano transcription system by 10% absolute by increasing training data diversity without requiring additional transcription. Automatically-labeled training examples ("semi-supervised" learning) gave a small improvement, but the biggest gains came from reusing our recordings for which ground-truth was available and detuning them by fractions of a semitone to simulate training from multiple pianos with slight tuning differences [7].

- The use of high-resolution frequency estimation based on instantaneous frequency is able to estimate the pitch of sung notes at the accuracy required for musicological investigations of intonation. By comparing the pitch tracking results for isolated voices with the results from an ensemble mixdown with added reverberation, we can quantify the impact of these complications: The ensemble recording achieves 90% of pitch estimates to within 0.2 semitones of those obtained from isolated voices; adding reverberation with a 0.9 sec RT_{60} reduced this to 75% of frames [1].

- A web-based music labeling game can be successful at gathering sufficient data to be a basis for training automatic classifiers to assign tags to music audio snippets, giving classifiers that correctly assign tags between 60% and 90% of the time, depending on the tag, on a balanced test set where random guessing gives 50% [5]. Although the ideas of multiple-instance learning, which considers the problem of knowing labels only for sets of instances without knowing how they apply to each instance individually, seem relevant, we have not yet been able to show any significant improvement over a naive approach of assuming the labels apply to all instances using these techniques [6].

- Our beat-synchronous chroma representation is a suitable foundation for automatic chord recognition, and we were able to quickly assemble a simple chord transcription system with performance equivalent to the state of the art [3].

- The chroma representation also has some benefit as a basis for automatically classifying music audio by artist (33% correct in a 1-of-20 task). Although

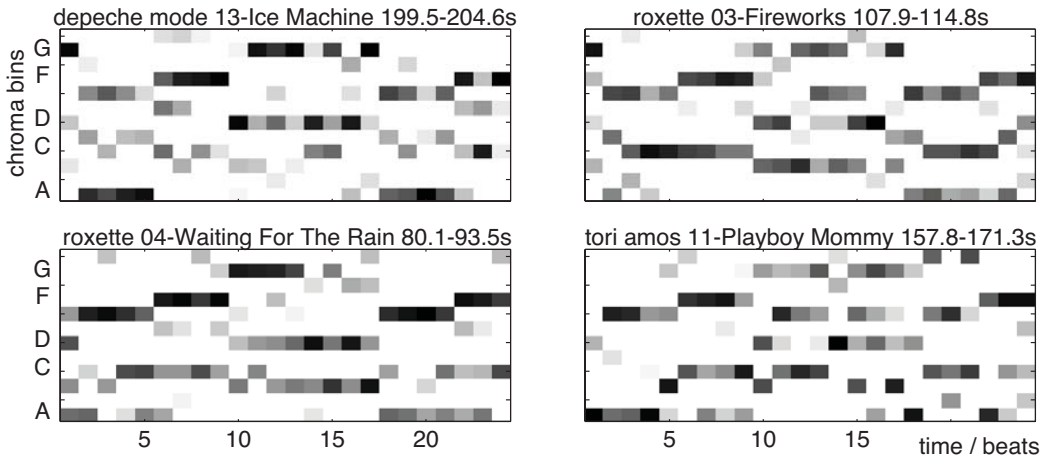

Figure 2: Example of similar beat-synchronous chroma fragments identified among the 1,413 tracks of the "artist20" dataset.

using the more common 'timbral' features such as MFCCs gives better performance in isolation (56% correct), combining timbral with chroma features gives classification that exceeds either feature alone (59% correct), implying that individual artists have some signature melodic-harmonic gestures that are being captured by this representation [2].

- The beat-synchronous chroma representation has provided a musically-rich yet computationally-efficient basis for matching and mining in large databases of music. Used as a basis for predicting human judgments of similarity between music clips, beat-chroma, which can capture melodic-harmonic resemblance while mostly invariant to timbre and instrumentation, perform far above a random baseline but not as good as conventional, timbral-based similarity, suggesting that in most cases it is timbral, rather than harmonic, similarity that listeners are judging [4]. Ongoing pilot investigations into finding recurring melodic-harmonic motifs in large collections of music audio show that Locality Sensitive Hashing is a viable mechanism for finding 'modes' in this high-dimensional space. Figure 2 gives an example of the results, where the hashing has identified four similar 24-beat phrases from four different songs, mostly in very different styles and instrumentation.

2 Training and development

This year, the project has provided support for graduate students Michael Mandel and Graham Poliner. Poliner graduated in May 2008, and a new graduate student, Graham Grindlay, has joined as of September 2008. The project has also provided material for a number of student projects supervised by PI Ellis, including one junior undergraduate, Vishal Kumar, who was supported over the summer while working on his music segmentation project.

3 Outreach

This year saw presentations by the PI and students from LabROSA at International Music Information Retreival Conference ISMIR (Vienna, Oct 2007), IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (Mohonk, Oct 2007), UCSD ECE department invited seminar (San Diego, November 2007), Digital Media Research Network +2 (London, December 2007), Centre for Interdisciplinary Research on Music, Mind, and Technology Distinguished Lecturer Serires, McGill University/University de Montréal (Montreal, March 2008), and IEEE ICASSP (Las Vegas, April 2008).

PI Ellis along with Jeff Snyder, doctoral candidate in Composition, participated in Siemens Science Day at Columbia on Saturday October 20th, 2007. This event for local high- and middle-school students, parents, and instructors attracted more than 1,300 participants, about 50 of whom attended the two sessions of the "Song Sights" interactive presentation on music and visualization.

PI Ellis completed co-editing a special issue of IEEE Tr. Audio, Speech, and Language, on the topic of Music Information Retrieval which appeared in February 2008.

LabROSA organized and hosted the first North East Music Information Special Interest Group meeting, a regional event to foster links among graduate students in this area, which attracted more than 40 participants to Columbia in January 2008 `http://labrosa.ee.columbia.edu/nemisig/`.

PI Ellis was general co-chair (with Youngmoo Kim of Drexel University) of the 2008 International Conference on Music Information Retrieval, ISMIR 2008, which had over 260 attendees and 105 full-length papers presented `http://ismir2008.ismir.net/`.

4 Contributions

4.1 Contributions to the principal discipline

- We have devised and developed the classification-based approach to music transcription, which is very different from the traditional approach of signal modeling, and offers high performance particularly in situations where good training data can be obtained.
- The Majorminer music labeling game has shown a practical way to collect listener tags closely associated with music audio, and we have demonstrated automatic audio tagging based on this data.
- We have introduced the idea of "multiple-instance learning" to the music audio analysis community and formulated music labeling problems in this framework.
- We have developed the beat-synchronous chroma representation, and expanded its application to domains of music similarity matching and classification. We have formulated a practical framework for identifying common musical motifs in large databases using this representation.

4.2 Contributions to sister disciplines

- Our collaboration with musicology student Johanna Devaney has allowed state-of-the art signal processing to be used to achieve high-resolution pitch tracking in polyphonic audio for the basis of analyzing vocal intonation practice, something that would not have been possible without these advanced techniques.
- Our collaboration with music composition students in the MEAP project has lead to many novel ideas, and is on the way to creating an interactive music-art installation that relies both on cutting-edge automatic music audio analysis and artistic creative vision.

4.3 Contributions to human resources

The project has provided direct support for two graduate students, summer support for one undergraduate, and has enabled the PI to participate in research supervision of several other students.

4.4 Contributions to educational infrastructure

The project has supported the continuing development of classroom and practical materials which we make freely available on our web site. These materials have been found valuable by many academics at multiple institutions worldwide.

4.5 Contributions to wider aspects of public welfare

Music is intimately valued by people from all areas of society. While our investigation is currently fairly abstract, our goal is to uncover deeper structure within music that will be relevant to every listener, and will give rise to generally-useful tools such as improved music discovery systems (to help listeners find music they can enjoy without relying on the marketing organizations of record companies) and novel music visualizations (to give insight into the way a piece of music 'works').

5 Resources made available

The following list includes relevant resources made available at PI Ellis's web site and elsewhere:

- MEAPsoft - software for analyzing, visualizing, and reordering music audio recordings: `http://www.meapsoft.org/`
- Automatic piano transcription: Data and code `http://labrosa.ee.columbia.edu/projects/piano/`
- Music Artist Identification: `artist20` Baseline System in Matlab (data and code) `http://labrosa.ee.columbia.edu/projects/artistid/`
- Artist Identification of Music Audio by Timbral and Chroma Features in Matlab `http://labrosa.ee.columbia.edu/projects/timbrechroma/`
- Class materials (slides, assignments, practicals, demonstrations) for Speech and Audio Processing and Recognition: `http://www.ee.columbia.edu/~dpwe/e6820/`
- Matlab examples of common audio processing algorithms: `http://www.ee.columbia.edu/~dpwe/resources/matlab/`
- MajorMiner: Music Labeling Game `http://www.majorminer.org/`

6 Publications

See references [1, 2, 3, 4, 5, 6, 7].

References

[1] J. Devaney and D. P. W. Ellis. An empirical approach to studying intonation tendencies in polyphonic vocal performances. *J. Interdisc. Music Studies*, 2(1–2):141–156, 2008. URL http://www.musicstudies.org/JIMS2008/articles/Devaney_JIMS_0821209.pdf.

[2] D. P. W. Ellis. Classifying music audio with timbral and chroma features. In *Proc. Int. Conf. on Music Info. Retr. ISMIR-07*, pages 339–340, Vienna, Austria, 2007. URL http://www.ee.columbia.edu/~dpwe/pubs/Ellis07-timbrechroma.pdf.

[3] D. P. W. Ellis. The 2008 labrosa audio chord detection system. In *MIREX 2008 System Abstracts*, 2008.

[4] D. P. W. Ellis, C. Cotton, and M. Mandel. Cross-correlation of beat-synchronous representations for music similarity. In *Proc. ICASSP*, pages 57–60, Las Vegas, Apr 2008.

[5] M. Mandel and D. Ellis. A web-based game for collecting music metadata. In *Proc. Int. Conf. on Music Info. Retr. ISMIR-07*, pages 365–366, Vienna, Austria, 2007. URL http://www.ee.columbia.edu/~dpwe/pubs/MandE07-majorminer.pdf.

[6] M. Mandel and D. P. W. Ellis. Multiple-instance learning for music information retrieval. In *Proc. ISMIR*, pages 577–582, Philadelphia PA, Sep 2008.

[7] G. Poliner and D. P. W. Ellis. Improving generalization for polyphonic piano transcription. In *Proc. IEEE Workshop on Apps. of Sig. Proc. to Audio and Acous.*, pages 86–89, Mohonk NY, Oct 2007. URL http://www.ee.columbia.edu/~dpwe/pubs/PoliE07-semisup.pdf.