EXTRACTING INFORMATION FROM MUSIC AUDIO

Information includes individual notes, tempo, beat, and other musical properties, along with listener preferences based on how the listener experiences music.

By Daniel P.W. Ellis

Music audio contains a great deal of information and emotional significance for human listeners. Machine systems that would be able to identify the listener-salient detail in music or that could predict listener judgments would be useful in many domains, from music theory to data compression to e-commerce. Here, I consider the kinds of information available in the music signal, reviewing current work in automatic music signal analysis, from the detection of individual notes to the prediction of listeners' music preferences. While this is a vibrant research area, including an annual international evaluation, researchers are still far from a working model of how listeners experience music.

Music is arguably the richest and most carefully constructed of all acoustic signals; several highly trained performers might work for hours to get the precise, desired effect in a particular recording. We can reasonably conclude that the amount of information carried by the musical waveform is greater than in any other sound, although it also gets us into the problematic territory of trying to define exactly what information it is that music carries, why it exists, and why so many people spend so much time creating and enjoying it.

Putting aside these philosophical points (they're beyond my scope here), we can name many objective aspects of a music recording (such as beat, melody, and lyrics) a listener might extract. As with other perceptual feats, we can hope to build computer-based systems to mimic these abilities. It will be interesting to see how well it can be done and consider the applications in which these systems might be used.

Music and computers have been linked since the earliest days of electronic computation, including the synthesis in 1967 by Max Matthews (then a researcher at Bell Labs) of "Daisy Daisy" on an IBM 7094 mainframe. Computer music synthesis soon led to the idea of computer music analysis, with the first attempt at automatic transcription in 1977 [9]. However, it was clear that, as with other attempts at machine perception, the seemingly effortless analysis performed by human senses were very difficult to duplicate on a machine. We are only now on the verge of having the algorithms, computational power, and data sets needed to produce systems that approach useful, general music transcription, along with various other musically relevant judgments. Meanwhile, technological developments have also presented urgent challenges in navigating large online and portable music collections that cry out for a "listening machine" able to hear, remember, and retrieve in listener-relevant terms.

Here, I look at a range of problems in extracting information from music recordings, starting with the most detailed (such as the individual notes played by a performer) and moving to high-level properties (such as musical genre applying to entire pieces or collections of recordings). However, the unifying theme is that abstract, symbolic information is extracted from raw audio waveforms. Thus, I do not include the significant body of work on making high-level musical inferences directly from score representations (such as machine-readable note-event descriptions like the Musical Instrument Digital Interface), though it has influenced more recent audio-based work.

Event-Scale Information

The information carried by music occurs at multiple levels, or timescales, each useful to automatic analysis systems for a variety of purposes. At the shortest timescale are the individual musical note events (such as indi-

vidual strikes on a piano keyboard). A musical score comprises a formal notation for these events, suitable for enabling a performer to play a piece. Music transcription is the process of recovering the musical score describing the individual notes played in a recording; we know it is possible because music students (after appropriate training) often do it very well. Transcription is valuable for searching for a particular melody within a database of recordings (needed for query by humming); high-quality transcripts would also make possible a range of analysis-resynthesis applications, including analyzing, modifying, and cleaning up famous archival recordings. A commercial example is Zenph Studios (www.zenph.com), a four-year-old startup that recreates damaged or noisy recordings of piano masterpieces by extracting the precise performance details, then re-rendering them on a robotic piano.

Musical pitch arises from local, regular repetition (periodicity) in the sound waveform, which in turn gives rise to regularly spaced sinusoid harmonics at integer multiples of the fundamental frequency in a spectral, or Fourier, analysis. Note that transcription could be a relatively simple search for a set of fundamental frequency Fourier components. However, such a search may be compromised for two main reasons:

Indistinctness. Noise, limitation of dynamic range, and the trade-off between time and frequency resolution makes identifying discrete harmonics in Fourier transforms unreliable and ambiguous; and

Interference. Simultaneous sinusoids of identical or close frequencies are difficult to separate, and conventional harmony guarantees that multiple-voice music is full of such collisions; even if their frequencies match, their relative phase may result in reinforcement or cancellation.

Nonetheless, many note transcription systems are based on fitting harmonic models to the signals and have steadily increased the detail extracted, ranging from one or two voices to higher-order polyphony. The range of acoustic conditions in which they can be applied has also increased, from small numbers of specific instruments to instrument-independent systems. Systems that transcribe notes from music audio include those described in [5, 6].

The Laboratory for the Recognition and Organization of Speech and Audio (LabROSA) at Columbia University has taken a more "ignorance-based" approach. There, my colleagues and I train general-purpose support-vector machine classifiers to recognize spectral slices (from the short-time Fourier transform magnitude, or spectrogram) containing particular notes based on labeled training data [3]. This data may be obtained from multitrack music recordings (each instrument in a separate channel), extracting the pitch of the main vocal line, then using the pitch values as labels for training features extracted from the full mixdown. This approach compares well to more traditional techniques, finishing third out of 10 systems in a 2005 formal evaluation of systems that identify the melody in popular music recordings. Conducted as part of the Music Information Retrieval Evaluation eXchange (MIREX-05) evaluations of music information retrieval technologies [2], it correctly transcribed approximately 70% of melody notes (on average). In many cases, transcribed melodies were clearly recognizable, implying transcripts are useful (such as for retrieval). But a significant number of excerpts had accuracies below 50% and barely recognizable transcripts. At LabROSA, our use of the classifier approach for detecting multiple simultaneous and overlapping notes in piano music has also worked well.

Individual note events may not be the most salient way to describe a piece of music, since it is often the overall effect of the notes that matters most to a listener. Simultaneous notes give rise to chords, and musical traditions typically develop rich structures and conventions based on chords and similar harmonies. Chords could be identified by transcribing notes, then deciding what chord they constitute, but it is easier and more robust to take a direct path of recognizing chords from the audio. The identity of a chord (such as C Major and E minor 7th) does not change if the notes move by multiples of one octave, so chord-recognition systems typically use so-called "chroma" feature instead of normal spectra. Where a spectrogram slice describes the energy in every distinct frequency band (10Hz–20Hz, 20Hz–30Hz, 30Hz–40Hz, and so on), a chroma feature collects all the spectral energy associated with a particular semitone in the musical scale (such as A) by summing the energy from all the octave transpositions of that note over some range (such as 110Hz–117Hz, 220Hz–233Hz, and 440Hz–466Hz).

Other chroma bins sum the energy from interleaved frequency combs. Since the combination of notes in a chord can produce a fairly complex pattern, chord-recognition systems almost always rely on trained classifiers; LabROSA borrows heavily from speech recognition technology, using the well-known expectation-maximization (EM) algorithm to construct hidden Markov models (HMMs). Each model describes a chord family, and the process of model estimation simultaneously estimates the alignment between a known chord sequence and an audio recording while avoiding the time-consuming step of manually mark-

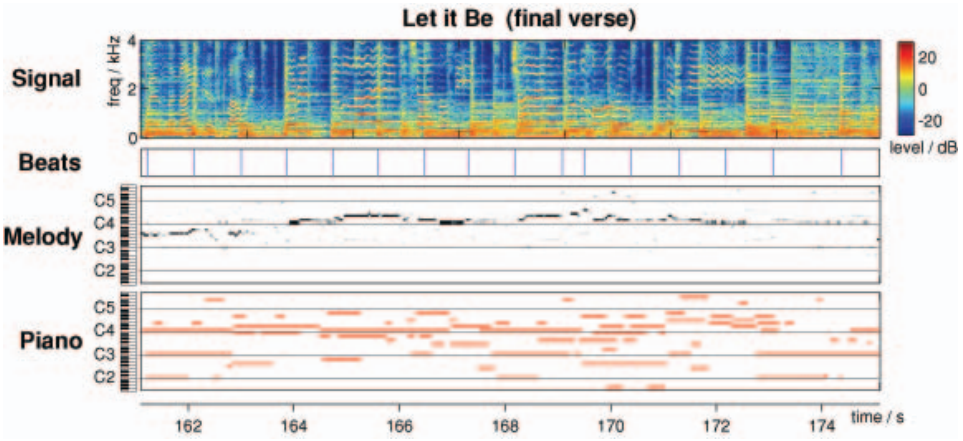

Figure 1. Example transcription for a fragment of "Let It Be" by the Beatles. Below a conventional narrowband spectrogram are automatically generated estimates of downbeat, melody note probability, and piano part. Notes are in a "piano-roll" representation, with horizontal stripes describing the activation (on the left) of adjacent notes, or keys on a piano.

ing the chord change times in training data [11].

Largely orthogonal to the note being played is the identity of the instrument playing the note. Percussion instruments have relatively little variation among note events and have been successfully identified (such as in pop music) for transcribing drum parts. However, despite considerable literature on recognizing the instrument in solo notes or phrases, recognizing one instrument in a mixture of instruments and voices is much more difficult—perhaps because statistical modeling techniques are overwhelmed by the huge variations in the accompanying instruments encountered.

Locating singing within pop music has been relatively successful, with several projects able to recover the precise temporal alignment between known lyric content and recordings. However, an unusual speaking style, along with significant non-speech energy, make direct transcription of lyrics, or speech recognition for songs, a significant challenge.

Figure 1 illustrates some of these approaches, showing a 14-second excerpt from a pop music recording ("Let It Be" by the Beatles) analyzed through various methods. The top pane shows the conventional narrow-band spectrogram in which note harmonics appear as horizontal ridges; drum sounds and other onsets appear as vertical stripes. Below that is a set of "downbeats," or the start of each beat unit, derived from a tempo-smoothed event detector. The next two panes show note sequences as a function of time in a "piano roll" format; each horizontal stripe corresponds to one semitone, or a key on a piano keyboard, and note events are indicated by dark stripes. The higher of the two panes shows a probabilistic estimate of the melody notes from the LabROSA melody extractor [3]; the lower pane estimates all notes played by the piano.

Phrase-Level Information

Individual notes are not music in and of themselves. Music emerges from the time-sequence of events, and a number of musically important properties can be recognized over multiple, successive note events, perhaps the individual phrases or lines that form the shortest recognizable musical fragments. Tempo is a basic property of a musical fragment. Although intuitively natural (such as the steady beat of foot-tapping) it is also intrinsically ambiguous, since much music seemingly "plays" with our perception of repetitive period. Approaches to tempo extraction from music audio can attempt to either extract note onsets, then fit periodicities to these event sequences, or use a mechanism (such as autocorrelation) to identify strong periodicities in the signal energy envelope without explicit event detection [10].

Automatic tempo extraction has obvious applications in playlist sequencing but is complicated by rhythmic ambiguities arising from the hierarchical structure of beats, with different tempos at different levels; for example, locking onto a quarter-note might give a tempo of 60 beats per minute (bpm), whereas identifying the eighth-notes would give 120 bpm. A related challenge involves finding the downbeat, or the start of each beat, as opposed to the spacing between beats.

Tempo extraction was another facet of the MIREX-05 evaluations, with many systems able to extract the tempo of at least part of the two-level "ground-truth" beat hierarchy. A "faster" and "slower" pulse are both consistent with the music, like the quarter-note/eighth-note ambiguity, in over 90% of a diverse set of musical excerpts. Finding the correct downbeat was much more difficult, with the best system correct in under 50% of the excerpts.

Musical phrases frequently extend beyond even the slowest tempo segmentation and have irregular lengths. Yet the repetition and alternation of segments (such as verse and chorus) are central to many kinds of music. Automatic identification of this structure can be valuable for music summarization; for instance, if a four-minute pop song is to be represented by a 10-second excerpt, it is probably best to use the first 10 seconds of the chorus, which can be defined as the most-repeated part of the song [7].

Several promising approaches to identifying music structure are based on the idea of identifying stretches of repetition by building a matrix of similarities between every pair of, say, 25-millisecond frames in the audio, then looking for diagonal ridges away from the leading diagonal. Such off-diagonal ridges indicate sequences of frames that precisely repeat one or more earlier segments [4]. Identifying these repetitions can lead to automatic labeling of verse, chorus, or other structural elements of the music. Figure 2 outlines an example of a similarity matrix derived from an entire pop song, with the human-labeled segment boundaries overlaid for comparison.

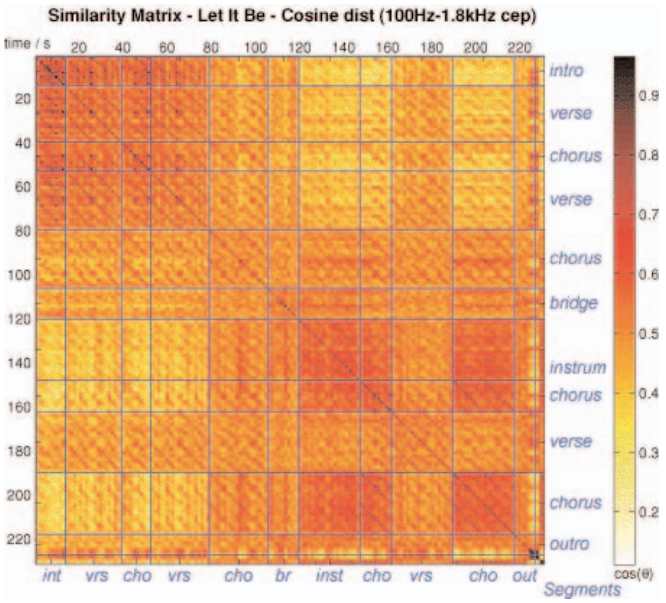

Figure 2. Example similarity matrix for "Let It Be." Each cell represents the similarity between 25-millisecond time frames extracted at different points in the music. Diagonal structures away from the leading diagonal indicate similar repetitions. Blue lines are human-labeled segment boundaries.

Piece-Level Information

The final level of information entails an entire piece or collection of pieces. Unlike the discussion so far in which I've concentrated on the "surface detail" of the music, describing particular listeners' music collections in order to recommend new artists requires a much more general and abstract description. Automatic music analysis systems must look at longer temporal extents, considering them from much broader (and hence typically less precise) categories.

Automatic labeling by music genres (such as "rock," "jazz," and "classical") [12] achieved 60% accuracy classifying pieces into 10 genres using the standard tools of statistical pattern recognition. Each piece was represented by short-term features (such as the Mel-Frequency Cepstrum Coefficients, or MFCC, commonly used in speech recognition to capture gross spectral structure within a few dimensions), as well as novel music-related statistics derived from a "pitch histogram" and "beat histogram." But genre classification can involve judgments like discriminating "rock" from "pop" over which even listeners often disagree; identifying the artist (or band) who created a particular piece has been used as a similarly abstract classification task but one for which the correct labels, or "ground truth," are easier to obtain and less subjective.

MIREX-05 involved large-scale tests of both genre and artist identification. LabROSA's system was tops in artist identification, correctly identifying which artist, from a list of 73 pop musicians, had created approximately 70% of 1,300 music audio files. The system [8] used only the covariance of MFCC features over the entire piece but a relatively powerful support vector machine classifier to infer decision surfaces around a class defined by arbitrary examples. These examples could equally well define artist, genre, or any other property.

Perhaps the most compelling application of music classification at the piece level involves automatic prediction of music preference for a particular user. Such a metric can be used for automatic playlist generation (by finding relevant pieces from a listener's existing collection) or for recommending new music (by searching a large archive for previously unknown pieces that resemble known, preferred songs). Existing commercial systems make such suggestions based on collaborative filtering; that is, they identify other listeners with similar music collections and recommend the novel components of their collections. However, basing recommendations on audio content, rather than on other listeners' behavior, removes bias toward well-known music and makes it just as easy to find interesting music from a range of obscure sources.

LabROSA has developed a browser that allows exploration by individual users of any large music collection by displaying pieces (or artists) in a neighborhood of similarity around a selection. The system models subjective similarity via an "anchor space" in which each dimension is the output of a classifier trained for a musically relevant judgment (such as "electronica" and "female singer"). Included are 12 such dimensions, and the browser allows direct manipulation of the current "location" to find music that is, say, "like the current artist but with a more country and western feel" [1].

Figure 3 is an example of the browser's front-end; devising objective evaluation standards is still a research topic. Although it seems unlikely that the low-level features would be able to capture much of the information relevant to high-level preferences, the experience of

Figure 3. The "playola" music similarity browser. The column on the left lists pieces by the currently selected artist. The column on the right lists the rating of that artist (or selected piece) in terms of 12 "anchor" dimensions (each shown as a separate bar), followed by a ranked list of similar pieces based on the anchor-space projection.

playing with such systems is sometimes startlingly precise.

Despite being able to correctly identify the genre or artist related to a given piece of music, the current best similarity judgment systems (rated at MIREX-05) do not derive demonstrable advantage from the detailed event- and phrase-level information described earlier. Such information is surely relevant to listeners, indicating considerable potential for improving the relevance and usefulness of these techniques.

Conclusion

Considering the range of information present in a music audio signal, audio researchers have made significant progress in the automatic extraction of musical attributes (such as pitch and instrumentation) at several scales, from individual notes up to entire collections of recordings. The quality of current extraction systems is, however, still generally far short of a trained music student or other human judge. This limitation is an obstacle to applications like ultra-low-bandwidth compression into event streams. Despite being woefully ignorant of the small but critical nuances of timing and tuning essential to musical beauty, extraction systems are still valuable in certain situations (such as query by humming and new music discovery) where the alternative would be for the user to wade through unmanageably large music collections. Progress in the field over the past decade and the exciting development of regular, formal, international evaluations over the past two years is a testament to the importance of music audio in our lives and to the value of automatic information extraction from these signals. C

References

1. Berenzweig, A., Ellis, D., and Lawrence, S. Anchor space for classification and similarity measurement of music. In *Proceedings of the IEEE International Conference on Multimedia and Expo ICME-03* (Baltimore, July 6–9, 2003), 29–32; www.ee.columbia.edu/~dpwe/pubs/icme03-anchor.pdf.
2. Downie, J., West, K., Ehmann, A., and Vincent, E. The 2005 Music Information Retrieval Evaluation eXchange (MIREX 2005): Preliminary overview. In *Proceedings of the Sixth International Symposium on Music Information Retrieval ISMIR* (London, Sept. 11–15). Queen Mary University of London, 2005, 320–323; www.music-ir.org/evaluation/mirex-results/.
3. Ellis, D. and Poliner, G. Classification-based melody transcription. *Machine Learning Journal* (2006); www.kluweronline.com/issn/0885-6125.
4. Foote, J. Visualizing music and audio using self-similarity. In *Proceedings of the ACM Multimedia Conference* (Orlando, FL, Oct. 30–Nov. 5). ACM Press, New York, 1999, 77–80; www.fxpal.com/publications/FXPAL-PR-99-093.pdf.
5. Goto, M. A real-time music scene description system: Predominant-F0 estimation for detecting melody and bass lines in real-world audio signals. *Speech Communication 43,* 4 (Sept. 2004), 311–329.
6. Klapuri, A. Multiple fundamental frequency estimation by harmonicity and spectral smoothness. *IEEE Trans. Speech and Audio Processing 11,* 6 (Nov. 2003), 804–816; www.cs.tut.fi/sgn/arg/klap/multiplef0.pdf.
7. Logan, B. and Chu, S. Music summarization using key phrases. In *Proceedings of ICASSP* (Istanbul, June 5–9, 2000), 749–752; citeseer.ist.psu.edu/logan00music.html.
8. Mandel, M., Poliner, G., and Ellis, D. Support vector machine active learning for music retrieval. *Multimedia Systems* (2006); dx.doi.org/10.1007/s00530-006-0032-2.
9. Moorer, J. On the transcription of musical sound by computer. *Computer Music Journal 1,* 4 (Winter 1977), 32–38.
10. Scheirer, E. Tempo and beat analysis of acoustic musical signals. *J. Acoust. Soc. Am. 103,* 1 (Jan. 1998), 588–601; web.media.mit.edu/~eds/beat/.
11. Sheh, A. and Ellis, D. Chord segmentation and recognition using EM-trained hidden Markov models. In *Proceedings of the International Conference on Music Information Retrieval ISMIR03* (Baltimore, Oct. 26–30, 2003); ismir2003.ismir.net/presentations.html.
12. Tzanetakis, G. and Cook, P. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing 10,* 5 (July 2002), 293–302.

Daniel P.W. Ellis (dpwe@ee.columbia.edu) is an associate professor in the Department of Electrical Engineering at Columbia University, New York.