Data-Driven Music Audio Understanding

Daniel P.W. Ellis and Douglas I. Repetto, Columbia University

The essence of music is *structure*; the composer Edgar Varèse called it “organized sound”. For centuries, people have responded to their enjoyment of music by studying and describing that structure, yet there is still no satisfactory description of what music *is*, much less how or why it works. This project proposes to describe and explain musical structure using automatic signal analysis and machine learning. Starting from a database of many thousands of examples of music audio, computers will mine for recurrent structures and patterns at successively higher levels to create rich but compact descriptions of the realizations of music, to complement and extend human musicological insights from an objective perspective. In order to work with very large databases it is critical to start from audio, rather than notated forms, since this is the only ‘canonical form’ in which all music exists; by the same token, audio is the representation closest to the original music.

Apart from the abstract goal of a deeper understanding of music, these results may also cast light on the general techniques for representing complex information within the brain. More practically, automatic systems for identifying high-level characteristics of music audio can revolutionize the way that listeners find and organize music, since their interests can be described in terms of objective, yet relevant, properties of the music audio itself. Finally, the strong appeal of music to all people, but particularly to school-age students, presents an opportunity to broaden interest in science and technology by developing and deploying a set of classroom materials and other tools that middle- and high-school students and teachers can use to analyze and modify the music of their own choice.

Intellectual Merit: Despite the enormous amount of knowledge and study relating to specifics of music, there is very little work that can truly be said to start from the minimum of assumptions and attempt to define music simply in terms of whatever statistical regularities actually occur in a large body of real, polyphonic music. The work we propose will begin to construct such a definition. Moreover, the specific tools for exposing musically-relevant aspects of the signal will have many applications in more specific music-processing applications such as automatic accompaniment or computer-aided composition, and the approaches to learning high-level, hierarchic structure may well have applications in other, analogous problems of identifying structure from large databases such as multimedia content analysis or natural language processing.

Broader Impact: Because music is important in the lives of so many people, new technologies to help give insight into the structure and ‘function’ of music have very broad potential impact. The project highlights the commercial possibilities of a system for music recommendation based on an analysis of a listener’s existing collection in terms of high-level attributes identified in the analysis (similar to the Pandora music recommendation service, but without requiring human experts to classify each piece). A significant part of the project is a plan for outreach to school-age students, building on novel interactive music analysis and resynthesis tools, and tied in to an existing summer school and public school program operated by Columbia University.

Keywords: Music audio, structure discovery, machine learning, signal processing.

Data-Driven Music Audio Understanding

Daniel P.W. Ellis and Douglas I. Repetto, Columbia University

1 Introduction

> *"The incompleteness of any existing formal descriptive theory of music implies that the only way that you can explain to another person the meaning of the word "music" is by example. In other words, if you were teaching some person English, and you didn't know any word for "music" in their language, you would be forced to play some music to them, and then tell them that that's what music is."*
> Philip Dorrell, What is music? Solving a scientific mystery [Dorrell, 2006].

Music is unparalleled in the emotions and attachment it can evoke in listeners, yet it is practically impossible to agree on a definition of what music *is* (other than indirectly through listeners' behavior), and certainly any accounts of *why* music 'works', or what distinguishes different kinds of music, are typically incomplete, inadequate, and suspect.

This project proposes a radical approach to defining and understanding music. Rather than examining expert insight into the features of particular kinds of music, we will use the objective tools of signal processing and machine learning to mine for common, repeating structure across a large collection of real music audio recordings. This analysis attempts to escape our preconceptions concerning the identity or location of music's essence, and to generalize the definition of musical form and structure by identifying the regularities that occur within phrases, pieces, the output of a particular musician, or an entire genre.

The explosion in digital music makes this project possible; whereas a decade ago a statistical analysis of 100 pieces would be considered large-scale, it is now not only feasible but relatively straightforward to consider analyzing 100,000 recordings in a broad category such as contemporary pop music. Such a large database comprises a powerful, but latent, definition of music; our goal is to uncover that definition and make it at least partially explicit.

Although interesting as a purely intellectual challenge, we motivate this work from two angles: a range of commercial opportunities that are made possible by better machine understanding of music, and the idea of leveraging the appeal of music to encourage the interest of school-age students (and others) in science, technology, and mathematics.

1.1 Commercial Applications: Predicting Musical Preference

Music is a multi-billion-dollar business that is rapidly being transformed by the influence of information technology. Novel automatic analyses of music audio content have enormous potential to influence this industry, something that the European Union have acknowledged through their substantial commitment to industry-oriented music analysis projects under the 5th Framework research program (e.g. SIMAC [Serra, 2004], SemanticHiFi [Vinet, 2003], S2S^2[Bernadini, 2004]).

One revolutionary visions is for a future in which the big record labels no longer act as the intermediaries between musicians and listeners, but instead anyone can obtain music matching their particular tastes directly from individual amateur or semi-professional musicians. This shift is a kind of musical equivalent to the way that blogs have supplanted the centralized monopoly newspapers and magazines in connecting readers to journalists – not necessarily complete replacement, but a dramatic broadening of choices and channels. The problem with applying this model to music is that it can be difficult and time-consuming to find any music that actually appeals to a specific listener's taste among the hundreds of thousands of musical offerings in the catalogues of online retailers, or made freely available on web sites such as MySpace.

Over the past couple of years, there has been a rush to provide a technological solution to this overload in music choice. Perhaps the most successful service to emerge is Pandora.com [Pandora, 2005], a kind of personalized online radio station, which uses a few examples of music preferred by a particular listener to choose 'similar' music to play, refining its choices based on feedback when the listener dislikes any individual track. Pandora is based on a database of several hundred thousand tracks tagged with around 400 high-level musical attributes such as "Piano playing", "Disco Influences", or "Narrative Lyrics". Listener preference is determined as the set of absent or present attributes common to all preferred music, and new tracks are chosen to fit this profile. Critically, these attributes are hand-defined and manually assigned to each piece; Pandora employs a team of several dozen music experts to label new music, and reports that it can take up to half an hour to complete the manual analysis of each piece. The net result, however, is a system that can recommend music largely independent of preconceived genre categories, and, unlike "listeners who liked this also liked..." collaborative filtering systems, one that is not biased by the current popularity of a particular piece or artist – provided that artist meets the minimum threshold of having been analyzed by the team of experts, which is clearly an expensive and limited resource.

Within this project we would like to provide a basis for duplicating the valuable and successful attribute-based Pandora system, but based entirely on objective analysis of the music – both in terms of the definitions of the attributes by which the music is described, and in the assignment of those attributes to particular music pieces. While using human judges is a clever and appropriate solution for rapid deployment of a successful system, particularly when the universe of music is largely limited to current commercial catalogs, it would be very interesting to see which automatically-derived high-level properties correlate best with user's preferences, and of course a fully automated analysis is required for an approach that can scale to and be sustained across the world of amateur and semi-professional musicians made possible by current low-cost recording and production technology. As detailed in section 2, we have been researching automatic music recommendation for several years, but a deeper, higher-level musical analysis appears to be the key missing ingredient to provide more useful and convincing recommendations.

1.2 Educational Applications: Outreach via Music Audio Analysis

Science and technology education can have a problem reaching out to the full community of students at the middle school and high school levels, where it may be perceived as dry and irrelevant. Significantly, this is also a period in the lives of young people when music – particularly the contemporary youth-culture music – can become extremely important. Thus, curriculum modules that introduce mathematical analysis and abstract structure by directly involving this music as the object of study provide a powerful vehicle to introduce technical concepts to students in a way that is exciting and relevant. Regardless of the wishes or intentions of individual musicians, all popular music from punk rock to disco to rap exhibits the same kinds of acoustic and formal structure that we aim to identify in this project – because that is the essence of what makes it music.

As described later in this proposal, the school of Engineering at Columbia already has extensive and growing involvement in public education in the surrounding Harlem community, and this project includes a substantial plan for transferring novel techniques and insights for music signal analysis into the classroom. By collaborating with local teachers we will develop our research nto specific tools and teaching modules to entice students with the power and potential of scientific, technological, and mathematical techniques. In many ways we see this as the most important aspect of our work.

2 Proposed Research

This section details the technical content of the project, introducing relevant prior work at each stage. Following this, in section 3, we will explain our plans for using these results to develop materials for school-age students.

2.1 General Overview: Structural Information from Music Audio

Music, captured as an audio recording, has rich meaning for a human listener but is largely opaque to computer systems – in the worst case, just a very long sequence of numbers describing the audio waveform. We have been developing techniques to obtain more flexible information from the audio, and figure 1 (modified from Ellis [2006a]) illustrates several of these, including melody extraction, piano transcription, tempo and beat tracking, and beat-synchronous chroma representation (for tempo- and instrumentation-independent representation of musical content).

The goal of this project is to form an automatic analysis of the structure inherent in music audio over a wide range of scales driven solely by the music audio itself. The key idea here is the automatic identification of approximate repetition – of a particular sound event, in a musical motif, within a piece, or across an entire archive of music. Given a large enough collection of examples, and algorithms that can take advantage of larger and larger amounts of data to create increasingly sophisticated analyses, it will be possible to derive a definition for music in terms of how it is actually encountered.

The idea of describing music in terms of individual notes is very familiar. Our hope, however, is to uncover levels of description at a higher level, for instance sets of notes or harmonic progressions that occur frequently within the music data. Given a dictionary of such common themes, a large amount of music could be compactly described, and as the dictionary captured the recurrent structure more effectively the description would become still more efficient. Moreover, descriptions in terms of dictionary elements at one level will themselves form patterns that can be analyzed and encoded at a higher level – patterns that may not even consist of the same elements, but which may constitute analogous structures in which the same patterns are realized using different elements. While this kind of music structure discovery will rely on the development of effective and sufficient low-level and mid-level representations, we believe that our work, combined with

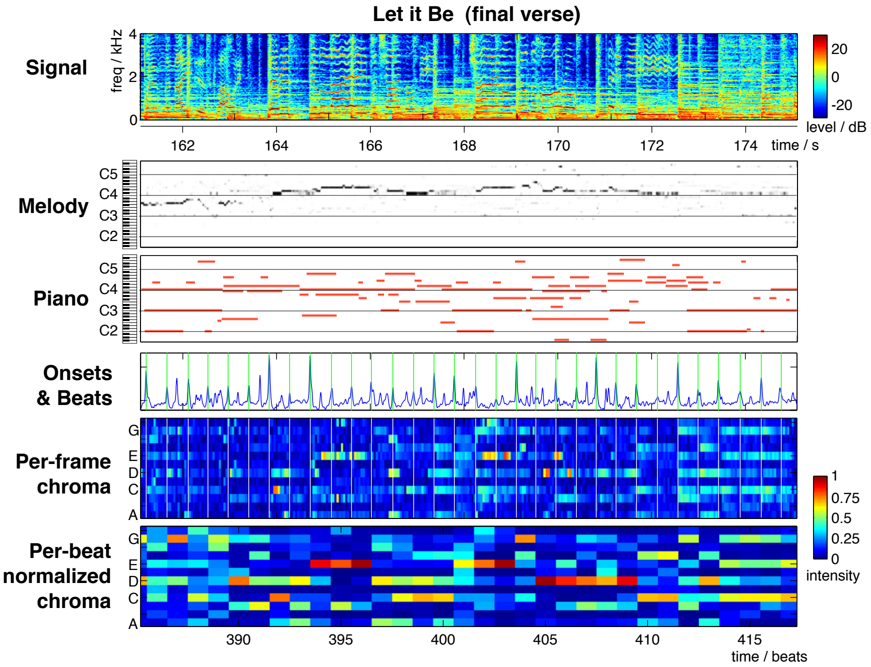


Figure 1: Extracting information from music audio. Top pane shows the spectrogram, the best-known tool for visualizing audio. Notes and other events are visible, but are not easily extracted by computer. Next panes show melody and piano transcriptions by the classifier systems described in the text. Below them are the beat track and per-frame and per-beat chroma features developed for music matching, also described in the text.

the unprecedented opportunity to process enormous music databases, will make this task feasible. The resulting insights into musical structure will then support both the commercial applications introduced above, and the educational goals of linking an emotional attachment to music with an enthusiasm for science and technology.

2.2 Low-level structure

2.2.1 Prior work: Note and chord transcription

By low-level musical structure we refer to the notes and individual sound events that are typically used to specify music by composers and performers. Trained musicians can transcribe music audio into notes with high accuracy, and researchers have been working on computer-based systems for several decades, starting with the work of Moorer [1975]. Our own work in **note transcription** (underlying part of the analysis shown in figure 1) achieves state-of-the-art performance while taking an unusual, data-driven approach, described below.

Most work in note transcription is based on the fact that a musical pitch is almost always associated with a near-periodic waveform, whose fundamental frequency corresponds to the pitch (e.g. 440 Hz for “Concert A”). Under a narrowband Fourier analysis, periodicity of this kind appears as a set of distinct harmonics at integer multiples of the fundamental frequency. By searching for this kind of pattern (e.g. in short-time Fourier transform magnitude peaks), the notes being played in a piece of music can in principle be deduced. To first order, multiple simultaneous notes lead to multiple sets of harmonics (although the nature of musical harmony means that harmonic collisions are more common that would occur at random). Systems that work along these lines include the work of Klapuri [2003] and Goto [2004]. They involve intricate and delicate systems for avoiding problems like harmonic overlap and octave errors.

We have investigated the idea of avoiding this expert design through the use of training data and machine learning [Ellis and Poliner, 2006]. Using multi-track music recordings (so that simple monophonic pitch trackers can create the ground truth for the final mix) and by synthesizing audio scores encoded as MIDI (a format designed for controlling music synthesizers) we created a set of labeled examples. We then trained simple but powerful support vector machine (SVM) classifiers on features such as the normalized Fourier magnitude spectrum and allowed the algorithm to learn which aspects of those features most reliably indicated the presence or absence of different notes. This has the nice property that second-order effects such as likely accompanying notes, and discriminating between near-miss cases like octave errors, are automatically included. Our system for melody transcription performed as well as the other top systems (within statistical margins of error) on an evaluation we helped organize as part of the 2005 Music Information Retrieval evaluation (MIREX; see section 2.5.1), reporting the correct melody pitch in around 67% of pitched frames [Poliner et al., 2007]. We trained similar classifiers to detect every note on the piano scale, and achieved a similar frame-level accuracy for recovering *all* notes being played, not just the melody line [Poliner and Ellis, 2006].

While notes are the obvious goal for low-level automatic transcription, they are not necessarily the best choice; we have also investigated direct **chord transcription**, where chords are the harmonic effect of note combinations. We took published chord sequences for a corpus of 20 songs by the Beatles and used an approach directly analogous to the way in which word-level transcripts of speech material are used to train speech recognition systems: Initial coarse alignments were successively refined using Expectation-Maximization (EM) until the alignments (and chord models) converged [Sheh and Ellis, 2003], avoiding the need for manual labeling. This approach has been validated by subsequent work in the area [Maddage et al., 2004, Bello and Pickens, 2005].

2.2.2 Proposed work: Data-driven sound atoms

While the work above learns the signal features directly from music examples, it requires the researcher to define the categories (notes and chords) in advance. The spirit of this proposal is to minimize or eliminate prior assumptions and learn as much as possible from the data alone. At the level of individual acoustic events, some recently-developed approaches that can infer elementary sound components based solely on their repeated appearance. The basic idea is that an audio waveform that consists of multiple instances of the same basic sound event – a particular note, or an instrument, or a percussion sound – can be efficiently described with a single ‘template’ instance of each event, associated with a set of time instants at which that template occurs (and perhaps one or two control parameters such as overall level for each occurrence). Given an algorithm that iteratively improves such a decomposition, for a fixed descriptive power (e.g. total number of templates) the algorithm should converge to a representation in terms of the actual distinct note events in the piece as the most efficient decomposition.

Non-negative matrix factorization (NMF) is such an algorithm [Lee and Seung, 1999]. In NMF, an inherently non-negative data set such as the short-time Fourier transform magnitudes (the spectrogram) is approximated as the outer product of two matrices,

$$\mathbf{X} = \mathbf{W}\mathbf{H} \tag{1}$$

where $\mathbf{X}$ is the spectrogram matrix with each column representing the spectral magnitude in a particular time window, and each row is the time-varying energy in one frequency band. The columns of $\mathbf{W}$ form a set of spectral templates, and the rows of $\mathbf{H}$ give the amplitude modulation associated with that template (which could be mostly zero for a sound occuring only a few times). Thus, the matrix equation 1 can be written as a sum across the K separate templates,

$$|X(t, f)| = \sum_{k=1}^{K} w_k(f) h_k(t) \tag{2}$$

where $w_k(f)$ is the k^{th} spectral envelope and $h_k(t)$ is its corresponding temporal envelope.

The surprising result is that with multiplicative update rules and appropriate normalization at each step, a locally-optimal solution to eqn. 1 can be found efficiently by gradient descent. The first application of this idea to spectrograms appears in Smaragdis and Brown [2003], but it has been followed by a great deal of work investigating this technique. A key development, proposed simultaneously by Smaragdis [2004] and Virtanen [2004], is to concatenate several successive spectra into a single super-column, allowing $\mathbf{W}$ to define not just spectral slices but entire time-frequency patches. Further, Schmidt and Mørup [2006] use a logarithmic frequency axis, where changes in fundamental frequency are approximately by shifts up and down the frequency axis, so that a single time-frequency patch in a column of $\mathbf{W}$ can be placed anywhere in time and at any transposition in frequency with the different elements of the corresponding row in $\mathbf{H}$.

A similar effect is achieved by building a dictionary based on recordings of individual instruments, then searching for combinations that fit an observation. In general, the dictionary will be overcomplete with multiple ways to achieve any observation. But a unique decomposition can still be found by including a sparseness criterion i.e. finding the best description that uses the smallest number of dictionary elements; While optimal sparse solutions are exponentially hard to find, the greedy matching pursuit algorithm gives good results and has been shown to go some way towards identifying and separating the notes of overlapping instruments [Leveau et al., 2006].

These approaches are usually applied to small fragments or single pieces. We are interested in scaling them up to create a single dictionary able to explain an entire corpus of music audio. Interestingly, sharing dictionary elements between pieces should both improve the efficiency gain and improve the quality of the elements themselves as they become based on increasingly large numbers of examples. Finally, knowing that two songs draw on similar or overlapping sets of elements may point to an important structural relationship.

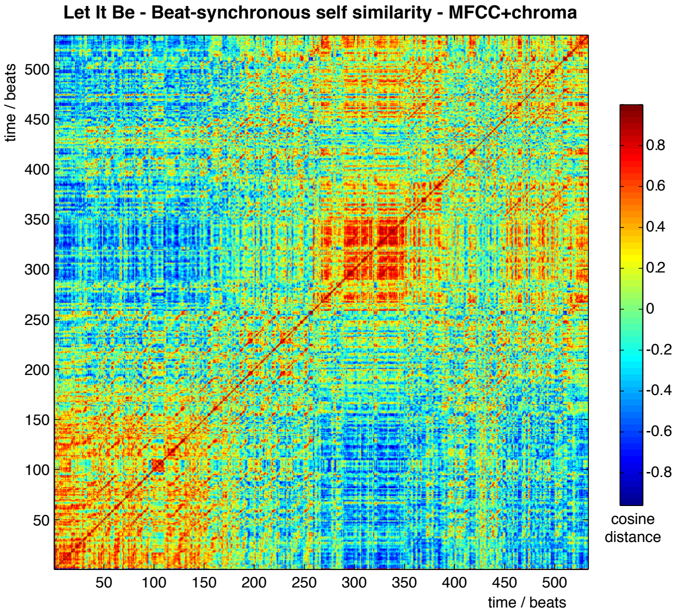


Figure 2: Mid-level structure within “Let It Be” by the Beatles. The image shows the normalized inner product (cosine distance) between feature vectors describing each beat-length segment of the song. The features are a concatenation of spectral (MFCC) and tonal (chroma) descriptors.

2.3 Mid-level structure

2.3.1 Prior work: Self-similarity and beat-synchronous features

Above the level of notes and individual events, music exhibits a range of structure within any given piece including metrical divisions (e.g. bars), phrases, and repeating segments such as the verse and chorus of conventional popular music. Detecting and revealing this kind of structure is a particularly rewarding way to connect a listener’s latent intuition about the structure of a piece of music with amore explicit representation. For example, figure 2 shows the ‘similarity matrix’ [Foote, 1999] – a full comparison of every time frame with every other time frame – for a beat-synchronous representation of Let It Be by the Beatles. The starting section, consisting of only piano and voice, lasts through beat 160 (i.e. 5 repetitions of the basic 32-beat phrase), whereupon the other instruments join, giving a different flavor (and making this region distinct). Subsequent shifts in the music are immediately visible.

Also visible in the self-similarity matrix are diagonal stripes away from the leading diagonal, indicating sets of frames with high similarity at a constant time difference. This is the signature of near-literal repeats of whatever aspects are reflected by the particular features being used, and have been employed to identify important recurrent sections in music e.g. to identify the chorus for music thumbnails [Bartsch and Wakefield, 2001, Goto, 2003].

Our contribution has been in the improvement of features, such as the beat-synchronous chroma features shown at the bottom of figure 1 and used as the basis for the similarity matrix in figure 2. Chroma features represent a short-time spectrum as a 12-dimensional vector where each bin aims to reflect the intensity of one particular semitone in a musical octave; they also naturally support transposition (the same tune in a different key) as a rotation. By reassigning spectral energy by semitone regardless of octave, chroma features can help normalize away variations in instrumentation etc. We use instantaneous frequency (IF) to obtain the sub-bin frequency resolution required to distinguish semitones at low pitches, and also to retain only strong, tonal peaks on the basis of consistent IF in adjacent frequency bins [Ellis and Poliner, 2007].

At the same time, we track beats in the music by forming an 'onset strength' waveform as the rectified first-order difference of log-magnitude auditory (mel) spectrum, then autocorrelating to find the strongest global tempo (weighted to fall into the range of tempos tapped by listeners), and using dynamic programming to find the best sequence of beat onsets that both reflect the target tempo and mostly line up with moments of high onset strength. This dynamic programming approach is both much simpler and generally more accurate than previous approaches, as shown by our joint-top performance in the 2006 MIREX beat tracking evaluation [Brossier et al., 2006] (MIREX is described in section 2.5.1).

Chroma features calculated at a fixed frame rate are averaged over beat-length segments to obtain a description for the musical piece which is reduced to 12 semitone bins by however many beats are detected by the beat tracker – a drastically smaller representation that still captures much of the melodic and harmonic essence of the piece, yet is largely invariant to tempo, instrumentation, etc. This representation is a good basis for identifying structure and repetition both within and between pieces, as discussed below.

2.3.2 Proposed work: Dictionaries of melodic-harmonic fragments

Our motivation in developing the beat-synchronous chroma features was to obtain a representation, applicable to a large database of music audio, that would allow us to discover a dictionary of fragments that can efficiently account for a large portion of data. These are the common themes, progressions, and riffs that may define the 'grammar' of a musical style. This problem is analogous to the decomposition into repeating atoms discussed in section 2.2.2, except that we do not need to be able to accommodate overlaps between fragments, so a simple scheme of segmenting into fixed-length units then applying k-means clustering, or a slightly more sophisticated approach of defining a set of dictionary templates, long enough to contain the largest fragment we expect to identify, then using EM to re-estimate the points in the database where they are used, as well as the means and variances of their values, will allow the data to determine the 'natural' length of each template by setting values outside that length to have large variances. Rotating chroma vectors will allow transposed versions of the same idiom to contribute to the same model.

Learning a dictionary across an entire corpus of, say, pop music (such as our standard 'uspop2002' collection [Ellis et al., 2003]), will give us one version of the 'definition of music' we envisioned in the introduction. In the same way that the vector-quantizer codebook of a low-bitrate speech codec attempts to cover all the possible variants of a speech spectrum while ignoring those that do not occur, a dictionary of chroma-sequence fragments will define a set of partial paths that occur in the database, and by implication exclude those that are not covered. While the absence of higher-level structure means this is not (yet) an interesting basis for algorithmically-generated music, it does provide a natural way to 'test' if a particular new example resembles the data used to train the dictionary – simply by measuring the distortion and/or data size resulting from attempting to represent the example with the dictionary. Dictionaries trained on the work of particular musicians could both provide an explicit list of their signature compositional moves, and provide a basis for comparing and discriminating between compositional styles.

Chroma features are not the only basis upon which these dictionaries could be learned, of course. Following on from the low-level work, we would prefer to learn patterns based solely on our data-derived atomic sound elements. An interesting challenge will be to establish 'equivalence classes' – sets of low-level elements that, though distinct, can be treated as the same for the purposes of identifying higher-level patterns, to allow a separation of, e.g., instrument and melody.

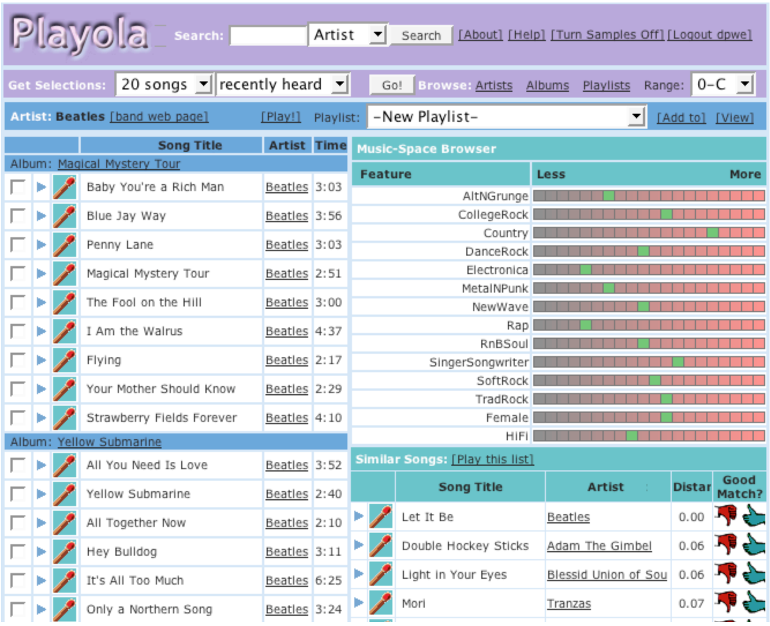
2.4 High-level structure

2.4.1 Prior work: Music similarity and matching

A high-level perspective might treat each separate recording as the smallest unit of consideration, then look for structure in the relationships between pieces, such as similarity over various dimensions, progressions, etc. We have conducted several projects concerned with music similarity and recommendation: Playola [Berenzweig et al., 2003a] is a music similarity browser that represents a piece of music as a distribution over a feature space, then allows the user to directly adjust different dimensions of that space to find 'neighboring' music. By taking the posterior probabilities of classifiers trained for different genres and other subclasses as the 12 'anchor' dimensions of the feature space, these adjustments have semantic interpretations such as "country", "new wave", "singer-songwriter", that provide some guidance for navigation.

Automatic measures of music similarity are a natural basis for playlist generation. We have built an automatic playlist generator that attempts to learn what the listener wants to hear via active learning. Starting with a single seed piece and some random alternatives, the system plays its next best guess of a similar song, whereupon the listener either approves the song, in which case it is added to the pool of positive examples, or skips (rejects) it, whereupon it is used as a counterexample [Mandel et al., 2006]. In use, this system appears remarkably successful at 'reading the mind' of the listener to quickly pick up on music of a particular artist or style. This SVM technique came top in the 2005 MIREX artist identification task [Mandel and Ellis, 2005].

While the previous systems were based on MFCC features originally developed for speech recognition and thus mainly reflect the instruments used in a piece, the beat-synchronous chroma features of section 2.3.1 provide a way to find similarity in harmonic-melodic structure largely independent of instruments. For the particular case in which different songs share long stretches of near-identical melody or harmony – e.g. cover versions, where a musician creates his or her own version of an existing work – direct comparison of the entire beat-chroma matrix can identify the matches. We built a system for cover song identification that simply cross-correlates the entire beat-chroma matrices of pairs of songs, at all 12 relative rotations of the chroma axis to capture key-shifted versions [Ellis and Poliner, 2007]. High-pass filtering the results to identify points where a particular time and chroma skew gave much higher correlation than immediate neighbors gives a relatively reliable basis for identifying covers; our system found more than twice as many true cover versions as the next best system in the MIREX-06 Audio Cover Song Identification evaluation [Downie et al., 2006].



Playola Search: Artist Search [About] [Help] [Turn Samples Off] [Logout dpwe]

Get Selections: 20 songs recently heard Go! Browse: Artists Albums Playlists Range: 0-C

Artist: **Beatles** [band web page] [Play!] Playlist: -New Playlist- [Add to] [View]

Song Title	Artist	Time
Album: Magical Mystery Tour		
Baby You're a Rich Man	Beatles	3:03
Blue Jay Way	Beatles	3:56
Penny Lane	Beatles	3:03
Magical Mystery Tour	Beatles	2:51
The Fool on the Hill	Beatles	3:00
I Am the Walrus	Beatles	4:37
Flying	Beatles	2:17
Your Mother Should Know	Beatles	2:29
Strawberry Fields Forever	Beatles	4:10
Album: Yellow Submarine		
All You Need Is Love	Beatles	3:52
Yellow Submarine	Beatles	2:40
All Together Now	Beatles	2:10
Hey Bulldog	Beatles	3:11
It's All Too Much	Beatles	6:25
Only a Northern Song	Beatles	3:24

Music-Space Browser

Feature	Less More
AltNGrunge	
CollegeRock	
Country	
DanceRock	
Electronica	
MetalNPunk	
NewWave	
Rap	
RnBSoul	
SingerSongwriter	
SoftRock	
TradRock	
Female	
HiFi	

Similar Songs: [Play this list]

Song Title	Artist	Distar	Good Match?
Let It Be	Beatles	0.00	
Double Hockey Sticks	Adam The Gimbel	0.06	
Light in Your Eyes	Blessid Union of Sou	0.06	
Mori	Tranzas	0.07	

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.

2.4.2 Proposed work: High-level structures and attributes

Starting from the harmonic-melodic fragments developed in the mid-level analysis, we can once again look for structure and repetition to identify high-order patterns and attributes. Modeling sequences of fragments, for instance with hidden Markov models (HMMs), offers the possibility of capturing longer-term structure that might actually be said to describe music, but, as with the relationship between low- and mid-level elements, our goal would be to find a way reuse high-level patterns of repetition with a variety of different elements ‘substituted in’, so that the high-level structure is describing only a general pattern of sequential repetition that can be rendered in a wide range of realizations.

By learning these patterns over specific subsets of music, or by clustering archive elements on the basis of the different structures that are matched or employed, we can begin to endow these data-derived entities with semantic attributes, for instance a song structure typical of pop, or jazz-influenced harmonic progressions. Given a large set of automatically obtained high-level attributes, the music collections of individual users can be modeled and clustered in the space of these dimensions, making it possible to identify which attributes are useful in predicting listeners’ musical tastes. It is not necessary to have explicit associations for each attribute, since the automatic analysis can use them anonymously. All that matters is that at least some of the attributes group together pieces in a way that corresponds to listener choices. Inspecting the automatically-derived structures that are frequently useful in explaining listener collections should reveal insights into what mediates musical taste.

2.5 Evaluation

Our broadest goal of defining music is not easily evaluated. However, we can define and perform evaluations of the more specific subtasks described above, and our practical goals of automatic music categorization and recommendation can also be evaluated quantitatively. In fact, many tasks of these kinds are already being rigorously evaluated under the recent MIREX initiative, which we describe briefly now.

2.5.1 MIREX

The International Conference on Music Information Retrieval, known as ISMIR, was held for the first time in 2000 and has occurred every year since, drawing hundreds of participants from academia and industry, and this year lasting four days and comprising 94 papers. Since 2004, the conference has also been associated with a formal evaluation of various aspects of automatic music analysis [Downie et al., 2005]. Our group has participated, and in many cases scored best, in several evaluations including artist and genre identification, melody extraction, beat tracking, and cover song identification; we actually organize the melody contest [Poliner et al., 2007] and we have contributed a substantial database for the artist and genre identification contests. The development of rigorous, respected, and well-supported evaluations has had a very positive effect on the field, giving specific answers to questions about the relative merits of different approaches which could not be addressed in the absence of common data sets, procedures, and metrics.

MIREX is run by volunteers, and each year anyone may propose any task for which a suitable test can be devised and which attracts a reasonable number of participants. As technologies in different areas become promising – for instance data-driven definition of sound events, identification of repeating structures within pieces, or discovery of high-level abstract characteristics shared within classes – we will propose these for inclusion within MIREX, which actually serves to encourage others to look at the related problems. These evaluations, in conjunction with the existing, relevant evaluations of transcription and music similarity, will provide a solid basis for measuring progress and demonstrating the value of our developments.

The high-level task of recommending new music to listeners based on the analysis of their preferences is one whose evaluation we have studied for several years [Ellis et al., 2002, Berenzweig et al., 2003b] including via direct, web-based surveys to gather subjective judgments of artist similarity. More recently, a large amount of behavioral information on actual listeners has been made available by Last.fm, a personalized online radio service that has several million users contributing the sequence of songs they play via a small reporting agent running on their computers. Last.fm makes this data freely available through an API [Last.fm, 2006], and this is an excellent basis for comparing different user preference models by measuring the likelihood of each successive track under the model of listener taste inferred from the tracks up to that point.

3 Education and Outreach

Because music has such power and appeal, this project has unusually rich opportunities for outreach, and the educational aspects are a central part of our proposal. We describe our plans in three categories: those belonging to traditional academic engineering activities, those arising from the interdisciplinary initiative between the Electrical Engineering and Music departments that was the impetus behind this proposal, and those that reach beyond the university to reach out to school-age students in our New York neighborhood.

3.1 Within-discipline activities

This proposal will support two graduate students through completion of their Ph.D.s in the area of music information extraction, which will prepare them for further research or employment in a wide range of multimedia content analysis and abstraction areas – topics which are undergoing an explosion in importance as web search moves beyond the confines of text documents. Their research will be disseminated by regular submission of papers to conferences and journals, as our group's strong publication record implies.

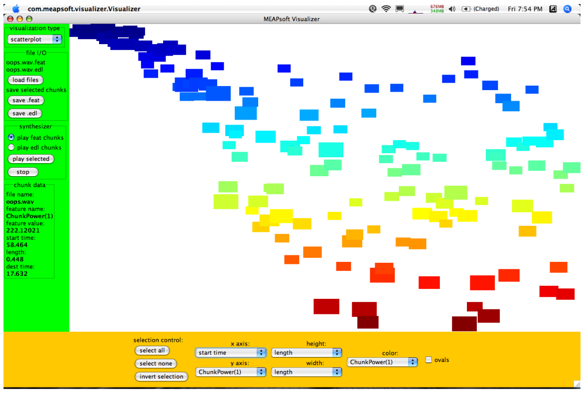
We will also use the results of the research, in conjunction with the earlier work on music information extraction reviewed above, to develop a new course on Music Information Processing Systems aimed at senior undergraduates and first-year graduate students. Currently, we include some of this material in two modules of a more advanced gradate course (Speech and Audio Processing and Recognition, ELEN E6820 [Ellis, 2006b]), but our department recently introduced a lower-level class on Music Signal Processing which has been very popular. A new course more oriented to abstract analysis than to studio processing will form a nice bridge.

As the amount of work in music information extraction grows, for instance as a result of relevant MIREX activities, we will organize a workshop to bring together interested researchers to share ideas. We have substantial experience organizing workshops on music and other topics [Ellis and Cooke, 2001, Divenyi et al., 2003, Raj et al., 2004, Divenyi et al., 2004, Eck et al., 2004, Raj et al., 2006] as well as related special issues [Cooke and Ellis, 2004, Ellis et al., 2006] and have found this to be a very valuable way to spark ideas advance a field.

3.2 Interdisciplinary activities

This project is proposed by a collaborative team from two departments at Columbia: Electrical Engineering and Computer Music Center (CMC). The PI and Co-PI have been collaborating for several years on a series of project classes – the Music Engineering Art Project, or MEAP [Ellis and Repetto, 2004] – that bring together students from the engineering and humanities disciplines in weekly meetings to share ideas and develop projects that express their varied interests spanning these areas; in all cases the students, though necessarily categorized as belonging to a particular department, are personally interested in topics spilling across these disciplines. A major motivation for these projects has been to address the kind of student who does not fit comfortably into existing academic compartments, and who we see as an increasingly common phenomenon in a world where technology is a tool as natural and widespread as basic literacy. MEAP has been made possible by seed funding from the university, but will need to secure external funding to continue beyond 2007.

Interdisciplinary collaboration is always an interesting challenge, and it has been particularly refreshing to take tools and techniques developed for tightly-defined engineering problems and



visualization type: scatterplot

file I/O: oops.wav.feat oops.wav.edl

save selected chunks

synthesizer: ● play feat chunks ○ play edl chunks

chunk data
file name: **oops.wav**
feature name: **ChunkPower(1)**
feature value: **222.12021**
start time: **58.464**
length: **0.448**
dest time: **17.632**

selection control:
x axis: start time
y axis: ChunkPower(1)
height: length
width: length
color: ChunkPower(1)
□ ovals

selection control:
x axis: dest time
y axis: ChunkPower(1)
height: length
width: length
color: ChunkPower(1)
□ ovals

Figure 4: MEAPsoft visualization: Each beat-length segment (in this case from the Britney Spears track "Oops I did it again") appears as a rectangle whose x and y co-ordinates are, in this case, given by start time and total power respectively. Left pane shows the original version, right shows the 'rearranged' version where a nearest-neighbor algorithm has been used to order the chunks..

look at them with a different perspective to see what broader uses they could have. The result of this, and the first substantial outcome of the MEAP project, is MEAPsoft, a program for rearranging music audio recordings [Weiss et al., 2006] [1] (see letter of support from Larry Polansky, Professor of Music, Dartmouth College). The MEAPsoft framework incorporates many of our research algorithms such as beat tracking, melody extraction, and similarity matching, to create a tool for music experimentation in which existing recordings are segmented at the beat level, then resequenced or interspersed based on a variety of principles, including sorting and similarity. The tool is continually growing, including advanced visualization capabilities as shown in figure 4, which are the inspiration for the ideas of revealing musical structure contained in this proposal. MEAPsoft has several unique capabilities, including an 'analogy' function in which the chunk sequence from one piece is used to order chunks with similar features (but most likely a very different order) from a second piece, creating an effective disentanglement of form and content.

3.3 Middle- and high-school outreach

To extend the scope of impact of the project beyond the university, we are proposing a three-step approach to achieve our goal of engaging high-school and middle-school students in the analysis of music and sound. The three steps are "Tools", "Teachers", and "Kids"; "Tools" is the basic research presented in section 2, and developed through the interdisciplinary collaboration described above. MEAPsoft is the first step towards packaging research algorithms in a format to make them available for a broader audience for experimentation and investigation. The same interdisciplinary team will develop more specifically school-oriented tools.

"Teachers" is the presentation of our work to knowledgeable science, technology, and music teachers from local schools. This step will take place within the context of the Engineering school's Center for Technology, Innovation, and Community Engagement (CTICE) run by Professor Jack McGourty [McGourty and Lowes, 2003] (see attached letter of support). This program,

[1] Freely downloadable at `http://labrosa.ee.columbia.edu/meapsoft/`.

supported by a GK12 grant from the NSF, provides curriculum materials and other support to science, technology, and math teachers in nearby public schools, which serve a community of students predominantly from typically underrepresented backgrounds. One of their regular events is a summer workshop for school teachers, and we have an invitation to present our tools in this workshop as a way to make contact with interested local teachers. By presenting to teachers we hope to learn from them how best to develop and orient such software for use with school students; our tools are powerful, but we recognize that they are not in the best form to inspire novice users to in-depth investigation. Collaborating with teachers will help us improve this accessibility, and find ways of talking about the material that will and engage middle- and high-school students.

The Tools and Teachers steps will form a feedback loop, and we will cycle through both steps several times in the course of the project. Things we learn from the teachers will be applied to the tools, and updates to the tools will in turn help us make further progress in curriculum development with the teachers.

The third step, "Kids", involves working directly with school students on this material, again through the CTICE activities. These include summer schools for both high-school and middle-school students; by the summer of year 2, we aim to have materials suitable for offering a project within this program. The second avenue for working directly with school students is by sending the graduate researchers supported by the project into local schools to work directly with our consulting teachers on delivering materials, following on from the CTICE GK12 Fellowship program. We will work with the students to help them understand the concepts involved, and the students will help us understand the effectiveness of our software and curriculum and how we might further refine them.

Our plans for this area are based in part on our relevant previous experience: From 2000-2002 the CMC led the "JPMorgan Chase Kids Digital Movement & Sound Project", a collaboration involving 15 middle school students from two local public schools in the creation of interactive sound and movement software and live performances. The project culminated with the entire group traveling to the third "World Summit on Media for Children" in Thessaloniki, Greece, where the students presented their performances to a large international audience. A CDROM consisting of the software as well as a movement+sound curriculum was distributed to schools around the world by JPMorgan Chase [Repetto et al., 2002].

We believe that compelling and engaging materials that show students how their love for music can be deepened via technological tools, combined with direct contact with Ph.D. students, will help to demystify academia and make science, math, and technology topics into attractive areas for further study.

4 Timetable

Year 1: *Research:* Development of basic data-driven sound-event vocabularies over large multi-song archives. Investigate approaches to within-song phrase-level segmentation using both local boundaries and repetitive structure. Develop mining for repeated motifs and gestures within 10,000 song database analyzed into beat-synchronous chroma features. *Education/Outreach:* Present available tools to schoolteachers at CTICE summer school. Initiate senior/graduate course on music information extraction.

Year 2: *Research:* Generalized structure discovery approaches at event, sequence, and collection levels. Develop better mechanisms and metrics for within-song structure decomposition *Education/Outreach:* Develop curriculum materials with schoolteachers, try in summer program for students. Organize structure discovery evaluation within MIREX.

Year 3: *Research:* Implement and evaluate high-level music similarity and preference attributes. Investigate other applications of abstract structural representation (compression, computer-supported composition, etc.). *Education/Outreach:* Experiment with using newly developed materials in schools. Organize workshop on music audio information extraction.

5 Summary

Music has a huge importance for listeners, and its impact and appeal present an irresistible intellectual mystery. While this project will not be able to answer “why” music works the way it does, it will create a novel and substantial answer to “what” music is by describing and abstracting the form and structure of music audio in large databases. This alternative means to musicological ends will, we believe, allow us to develop a set of materials, demonstrations, and tools to help all kinds of people find new and deeper insights into the music they love, and at the same time help ignite interest and excitement among school-age students into the potential of science, technology, and mathematics.

6 Results from prior support

PI Ellis is currently in the fourth year of NSF project IIS-0238301 “CAREER: The Listening Machine: Sound Source Organization for Multimedia Understanding” ($500,000, award period 2003-06-01 to 2008-05-31), concerned with analyzing sound mixtures by source. One student developed a “deformable spectrogram” model [Reyes-Gomez et al., 2005, 2004] and graduated in 2005. Current work includes segmentation and classification of personal audio recordings [Ellis and Lee, 2004b,a, 2006, Lee and Ellis, 2006].

Ellis is also in the first year of NSF project IIS-05-35168 “Separating Speech from Speech Noise”, a collaboration between Columbia, Ohio State, Boston University, and EBIRE ($750,000 total, $180,000 to Columbia, 2006-01-01 to 2008-12-31). which aims to understand how distortion and interference affect intelligibility, and thus improve automatic source separation for listeners. Preliminary results on our research in separation algorithms are reported in Ellis and Weiss [2006], Weiss and Ellis [2006].

Ellis was also a co-PI on NSF project IIS-0121396 “Mapping Meetings: Language Technology to make Sense of Human Interaction” ($1,402,851, award period: 2001-09-01 to 2005-08-31, PI: Nelson Morgan, ICSI). This project was concerned with the application of recognition techniques to information extraction from meetings. Our publications include Renals and Ellis [2003], Kennedy and Ellis [2003, 2004], Ellis and Liu [2004].

References

Mark A. Bartsch and Gregory H. Wakefield. To catch a chorus: Using chroma-based representations for audio thumbnailing. In *Proc. IEEE Worksh. on Apps. of Sig. Proc. to Acous. and Audio*, 2001. http://musen.engin.umich.edu/papers/bartsch_wakefield_waspaa01_final.pdf.

Juan P. Bello and Jeremy Pickens. A robust mid-level representation for harmonic content in music signals. In *Proc. Int. Conf. on Music Info. Retrieval (ISMIR)*, 2005. URL http://ismir2005.ismir.net/proceedings/1038.pdf.

Adam Berenzweig, Daniel P. W. Ellis, and Steve Lawrence. Anchor space for classification and similarity measurement of music. In *ICME 2003*, 2003a.

Adam Berenzweig, Beth Logan, Daniel P.W. Ellis, and Brian Whitman. A large-scale evalutation of acoustic and subjective music similarity measures. In *Proc. Int. Conf. on Music Info. Retrieval ISMIR-03*, 2003b.

Nicola Bernadini. S2S2: Sound to sense, sense to sound, 2004. URL http://www.s2s2.org/. European Commission FW6 project.

Paul M. Brossier, Matthew Davies, and Martin F. McKinney. Mirex-2006 audio beat tracking evaluation, 2006. URL http://www.music-ir.org/mirex2006/index.php/Audio_Beat_Tracking.

Martin Cooke and Daniel P. W. Ellis. Introduction to the special issue on the recognition and organization of real-world sound. *Speech Communication*, 43(4):273–274, 2004.

Pierre Divenyi, Daniel P. W. Ellis, and DeLiang Wang. Perspectives on speech separation: A workshop, November 2003. URL http://www.ebire.org/speechseparation/.

Pierre Divenyi, Nat Durlach, Daniel P. W. Ellis, and DeLiang Wang. Workshop on Speech separation and comprehension in complex acoustic environments, November 2004. URL http://labrosa.ee.columbia.edu/Montreal2004/.

Philip Dorrell. *What is music? Solving a scientific mystery.* 2006. URL http://whatismusic.info/.

J. Downie, K. West, A. Ehmann, and E. Vincent. The 2005 Music Information Retrieval Evaluation eXchange (MIREX 2005): Preliminary overview. In *Proc. International Conference on Music Information Retrieval ISMIR*, pages 320–323, London, 2005. URL http://www.music-ir.org/evaluation/mirex-results/.

J. Stephen Downie, Kris West, Elias Pampalk, and Paul Lamere. Mirex-2006 audio cover song identification evaluation, 2006. URL http://www.music-ir.org/mirex2006/index.php/Audio_Cover_Song.

Douglas Eck, D. P. W. Ellis, Ali Taylan Cemgil, and Jean-Francois Paiement. Workshop on Music Information Processing Systems (MIPS), December 2004. URL `http://www.iro.umontreal.ca/~eckdoug/mips/`.

D. P. W. Ellis. Extracting information from music audio. *Comm. Assoc. Comput. Mach.*, 49(8):32–37, Aug 2006a. URL `http://www.ee.columbia.edu/~dpwe/pubs/Ellis06-musicinfo-cacm.pdf`.

D. P. W. Ellis. Speech and audio processing and recognition, 2006b. URL `http://www.ee.columbia.edu/~dpwe/e6820/`. online course materials.

D. P. W. Ellis and M. Cooke. Workshop on Consistent and Reliable Acoustic Cues crac-01, August 2001. URL `http://www.ee.columbia.edu/crac/`.

D. P. W. Ellis and K. Lee. Minimal-impact audio-based personal archives. In *Proceedings of the 1st ACM Workshop on Continuous Archival and Retrieval of Personal Experiences*, New York, NY, October 2004a. URL `http://www.ee.columbia.edu/~dpwe/pubs/carpe04-minimpact.pdf`.

D. P. W. Ellis and K. Lee. Features for segmenting and classifying long-duration recordings of “personal” audio. In *Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing SAPA-04*, Jeju, Korea, October 2004b. `http://www.ee.columbia.edu/~dpwe/pubs/sapa04-persaud.pdf`.

D. P. W. Ellis and K. Lee. Accessing minimal-impact personal audio archives. *IEEE MultiMedia*, 13(4):30–38, Oct-Dec 2006. URL `http://www.ee.columbia.edu/~dpwe/pubs/EllisL06-persaud.pdf`.

D. P. W. Ellis and J. Liu. Speaker turn segmentation based on between-channel differences. In *Proc. NIST Meeting Recognition Workshop*, Montreal, March 2004. URL `http://www.ee.columbia.edu/~dpwe/pubs/nist04-turnid.pdf`.

D. P. W. Ellis and G. Poliner. Identifying cover songs with chroma features and dynamic programming beat tracking. In *Proc. ICASSP-07*, Hawai’i, 2007. URL `http://www.ee.columbia.edu/~dpwe/pubs/EllisP07-coversongs.pdf`. submitted.

D. P. W. Ellis and G. E. Poliner. Classification-based melody transcription. *Machine Learning*, Online First, 2006. URL `http://www.ee.columbia.edu/~dpwe/pubs/EllisP06-melody.pdf`.

D. P. W. Ellis and D. I. Repetto. Music Engineering Art Project (MEAP), 2004. URL `http://works.music.columbia.edu/MEAP/`.

D. P. W. Ellis and R. J. Weiss. Model-based monaural source separation using a vector-quantized phase-vocoder representation. In *Proc. ICASSP-06*, Toulouse, 2006. URL `http://www.ee.columbia.edu/~dpwe/pubs/`.

D. P. W. Ellis, B. Whitman, A. Berenzweig, and S. Lawrence. The quest for ground truth in musical artist similarity. In *Proc. Int. Symposium on Music Inform. Retriev. (ISMIR)*, pages 170–177, 2002. http://www.ee.columbia.edu/~madadam/papers/ellis02truth.pdf.

D. P. W. Ellis, A. L. Berenzweig, and B. Whitman. The“uspop2002” pop music data set, 2003. http://labrosa.ee.columbia.edu/projects/musicsim/uspop2002.html.

D. P. W. Ellis, B. Raj, J. C. Brown, M. Slaney, and P. Smaragdis. Editorial – special section on statistical and perceptual audio procesing. *IEEE Trans. Audio Speech and Lang. Proc.*, 14(1): 2–4, 2006.

J. Foote. Visualizing music and audio using self-similarity. In *Proc. ACM Multimedia*, pages 77–80, 1999. URL http://www.fxpal.com/PapersAndAbstracts/papers/foo99b.pdf.

M. Goto. 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):311–329, 2004.

Masataka Goto. A chorus-section detecting method for musical audio signals. In *Proc. ICASSP-2003*, pages V–437–440, Hong Kong, 2003.

L. Kennedy and D. P. W. Ellis. Pitch-based emphasis detection for characterization of meeting recordings. In *Proc. Automatic Speech Recognition and Understanding Workhop IEEE ASRU 2003*, December 2003.

L. Kennedy and D. P. W. Ellis. Laughter detection in meetings. In *Proc. NIST Meeting Recognition Workshop*, Montreal, March 2004. URL http://www.ee.columbia.edu/~dpwe/pubs/nist04-laughs.pdf.

A. Klapuri. Multiple fundamental frequency estimation by harmonicity and spectral smoothness. *IEEE Trans. Speech and Audio Processing*, 11(6):804–816, 2003. URL http://www.cs.tut.fi/sgn/arg/klap/multiplef0.pdf.

Last.fm. Audioscrobbler: The social music technology playground, 2006. URL http://www.audioscrobbler.net/data/webservices/.

D.D. Lee and H.S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

K. Lee and D. P. W. Ellis. Voice activity detection in personal audio recordings using autocorrelogram compensation. In *Proc. Interspeech*, Pittsburgh, PA, 2006. URL http://www.ee.columbia.edu/~dpwe/pubs/LeeE06-vad.pdf.

P. Leveau, E. Vincent, L. Daudet, and G. Richard. Mid-level sparse representations for timbre identification: design of an instrument-specific harmonic dictionary. In *Workshop on Learning the Semantics of Audio Signals*, Athens, Greece, 2006.

N. C. Maddage, C. Xu, M. S. Kankanhalli, and X. Shao. Content-based music structure analysis with applications to music semantics understanding. In *Proc. ACM Int. Conf. on Multimedia*, New York NY, 2004. URL `http://portal.acm.org/citation.cfm?id=1027527.1027549`.

M. 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. URL `http://www.ee.columbia.edu/~dpwe/pubs/ismir05-svm.pdf`.

M. I. Mandel, G. E. Poliner, and D. P. W. Ellis. Support vector machine active learning for music retrieval. *ACM Multimedia Systems Journal*, accepted for publication, 2006. URL `http://www.ee.columbia.edu/~dpwe/pubs/MandPE06-svm.pdf`.

Jack McGourty and Susan Lowes. Technology integration partnerships: Integrating new media technologies into teacher development, 2003. URL `http://tip.columbia.edu/`. NSF GK12 supported project.

James A. Moorer. *On the Segmentation and Analysis of Continuous Musical Sound by Digital Computer*. PhD thesis, Department of Music, Stanford University, 1975.

Pandora. Pandora media: The music genome project, 2005. URL `http://www.pandora.com/corporate/mgp`.

G. Poliner and D. P. W. Ellis. A discriminative model for polyphonic piano transcription. *EURASIP Journal on Applied Signal Processing*, 2006. URL `http://www.ee.columbia.edu/~dpwe/pubs/PoliE06-piano.pdf`. Special Issue on Music Signal Processing.

G. E. Poliner, D. P. W. Ellis, A. Ehmann, E. Gómez, S. Streich, and B. Ong. Melody transcription from music audio: Approaches and evaluation. *IEEE Tr. Audio, Speech, Lang. Proc.*, 2007. URL `http://www.ee.columbia.edu/~dpwe/pubs/PolEFGSO07-meleval.pdf`. accepted for publication.

Bhiksha Raj, Paris Smaragdis, and D. P. W. Ellis. Workshop on Statistical and Perceptual Audio Processing SAPA2004, October 2004. URL `http://www.sapa2004.org/`.

Bhiksha Raj, D. P. W. Ellis, Paris Smaragdis, and Malcolm Slaney. Workshop on Statistical and Perceptual Audio Processing SAPA2006, October 2006. URL `http://www.sapa2006.org/`.

Steve Renals and Daniel P.W. Ellis. Audio information access from meeting rooms. In *Proc. ICASSP-2003*, 2003. URL `http://www.dcs.shef.ac.uk/~sjr/pubs/2003/icassp03-mtg.html`.

Douglas Repetto, Thanassis Rikakis, Brad Garton, James Bradburne, Paul Kaiser, Dan Trueman, David Birchfield, R. Luke DuBois, Jason Freeman, Douglas Geers, and Kate Hofstetter. The JPMorgan Chase kids digital movement and sound project, 2002. URL `http://music.columbia.edu/kids`.

M. Reyes-Gomez, N. Jojic, and D. P. W. Ellis. Towards single-channel unsupervised source separation of speech mixtures: The layered harmonics/formants separation-tracking model. In *Proc. Workshop on Statistical and Perceptual Audio Proc. SAPA-04*, Jeju, Korea, October 2004. URL http://www.ee.columbia.edu/~dpwe/pubs/sapa04-transform.pdf.

M. Reyes-Gomez, N. Jojic, and D. P. W. Ellis. Deformable spectrograms. In *Proc. AI and Statistics*, Barbados, 2005. URL http://www.ee.columbia.edu/~dpwe/pubs/aistats05-defspec.pdf.

M. N. Schmidt and M. Mørup. Nonnegative matrix factor 2-D deconvolution for blind single channel source separation. In *ICA2006*, Charleston, SC, apr 2006. URL http://www2.imm.dtu.dk/pubdb/p.php?4061.

X. Serra. Semantic interaction with music audio contents (SIMAC), 2004. URL http://www.semanticaudio.org/. European Commission FW6 project.

Alexander Sheh and Daniel P.W. Ellis. Chord segmentation and recognition using em-trained hidden markov models. In *Proc. Int. Conf. on Music Info. Retrieval ISMIR-03*, 2003.

Paris Smaragdis. Non-negative matrix factor deconvolution: Extraction of multiple sound sources from monophonic inputs. In *Proc. Int. Congress on Independent Component Analysis and Blind Signal Separation*, volume 3195, page 494, 2004. ISBN 3-540-23056-4. URL http://www.merl.com/publications/TR2004-104/.

Paris Smaragdis and Judith C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *Proc. IEEE Workshop on Apps. of Sig. Proc. to Acous. and Audio*, Mohonk NY, 2003. URL http://www.ee.columbia.edu/~dpwe/e6820/papers/SmarB03-nmf.pdf.

Hugues Vinet. Semantic hi-fi: Browsing, listening, interacting, performing, sharing on future hifi systems, 2003. URL http://shf.ircam.fr/. European Commission FW6 project.

T. Virtanen. Separation of sound sources by convolutive sparse coding. In *Proc. ISCA Workshop on Statistical and Perceptual Audio Processing*, Jeju, Korea, 2004. URL http://www.sapa2004.org/papers/55.pdf.

R. J. Weiss and D. P. W. Ellis. Estimating single-channel source separation masks: Relevance vector machine classifiers vs. pitch-based masking. In *Proc. Workshop on Statistical and Perceptual Audition SAPA-06*, pages 31–36, Pittsburgh PA, 2006. URL http://www.sapa2006.org/papers/139.pdf.

Ron Weiss, Douglas Repetto, Mike Mandel, Dan Ellis, Victor Adan, and Jeff Snyder. Meapsoft: A program for rearranging music audio recordings, 2006. URL http://labrosa.ee.columbia.edu/meapsoft/.