Modeling Music Similarity: Automatic Prediction of Subjective Preference
Daniel P.W. Ellis, Electrical Engineering, Columbia University

Summary

Music evokes a powerful emotional response, yet it is very difficult to describe or explain why we like certain pieces of music and dislike others. As an example of a subjective judgment it is doubly obscure: firstly because it varies so much between individuals (each of whom is very sure about their own beliefs), and secondly because it is based, at least in part, upon our perception of the patterns and structure in music, itself a mysterious process.

In this program, we will investigate computational models of music similarity, specifically models that predict and account for subjective music preference in human listeners. It brings together three novel techniques: acoustic-based music structure discovery, automatic grounding of descriptive terms, and geometric models of user preference.

The research will draw upon acoustic signal separation and segmentation algorithms adapted from speech recognition, including EM model estimation, segmentation based on statistical significance tests, and forced alignment of training data to imperfect transcriptions. To relate these signal-related features to subjective qualities relevant to listeners, natural language processing techniques will be used to extract candidate descriptive terms from the text of automatically-mined web pages about music. A subset of these terms will be chosen for their discriminability (low information overlap) and their cohesion in the acoustic analysis space. The multidimensional space so defined, in which each musical recording is a point or a short trajectory, will be used as a domain to model the preferences of individual listeners, thus reflecting users’ differing sensibilities as different weightings for each dimension. This modeling will also incorporate the results of more formal subjective evaluations. Once a particular user’s preference has been modeled, a recording of a new piece of music can be analyzed, projected into the space, and rated according to her preference predictions.

Taken together, these modeling techniques could support an entirely new type of music listening, sifting through the work of many thousands of musicians who publish their music electronically outside the mainstream distribution channels. In this way, listeners can be connected to artists who match their tastes but who they might never have heard of otherwise, and individual musicians can be efficiently connected to their most enthusiastic audiences.

Thanks to the special allure of music, this project will have broader benefits in education and improving the public’s opinion of science and engineering: Much of the research is based on outstanding class projects done by students in my courses, and further results will feed back into the curriculum. The project will also provide more of the demonstrations that have proven so popular at engineering open days hosted by Columbia for local New York area students. Projects dealing with music are very media-friendly, and we anticipate popular coverage of our work, presenting academic research that is directly relevant and accessible to the general population.

Finally, the unprecedented span of an automatic analysis that connects raw signal properties to several levels of models of deeply personal reactions will establish a paradigm with broad significance in diverse fields including e-commerce, product design and usability, and machine perception.
1 Introduction: Musical Similarity and Preference

“There has to be some way for people to learn about new material. That’s what drove Napster. When Napster was at its peak, it was easier to find a lot of truly obscure music. […] There is so much potential for these technology companies to collect interesting data about what people are listening to and then make some intelligent recommendations. You could have a semipersonalized stream that would allow you to experience a radio station truly targeted to you.”

Shawn Fanning, creator of Napster, quoted in the New York Times Magazine. (Tapper 2002)

Music is unparalleled in its ability to evoke a powerful emotional response that is largely impossible to explain or inspect. Many people are passionately attached to the music they love, yet may be completely bemused by the tastes of their peers. This project seeks to investigate the nature of musical preference by seeing how well we can describe and predict individual taste using information derived principally from the recorded signal.

Analyzing musical preference is a fascinating challenge because it combines both enormous subjective variability with deep structural analysis. Preference for a favorite color is highly variable, but is for the most part based on a relatively straightforward photometric analysis of visual images. On the other hand, recognizing that a picture contains a Mercedes-Benz requires extremely sophisticated visual analysis and object detection, but is largely objective assuming the appropriate analysis is available (such as detection of the hood ornament). Predicting that a given listener will enjoy a particular piece of music combines both these difficulties, since we assume that preference is largely or at least significantly influenced by a high-level analysis in terms of harmony, instruments, and genre; yet even given perfect information in these terms, the question is still far from solved.

This fluidity is at the same time attractive from the point of view of the current problem. With few preconceptions about the answers we will find, we propose to use modern tools of machine learning in conjunction with large datasets made possible with current information technology (including mining the web) to search through large numbers of potential subjective characteristics, learning how well each one can be correlated with signal-derived structures, and in turn learning how well the preference patterns of individual users can be described in terms of these subjective qualities.

More specifically, this project will combine three hitherto separate research threads to create a system that spans an unprecedented range from signal to sentiment: Firstly, we will advance current work in structural analysis of musical signals (i.e. transcription of themes and harmonies) using data-driven machine-learning techniques that are unprecedented for these kinds of signals. Secondly, we will exploit ‘community metadata’ – such as text drawn from web pages discussing particular pieces of music – to define candidate subjective qualities that may be relevant to the music, then use modern machine-learning techniques such as Support Vector Machines to build
classifiers for these subjective qualities based on the high-level, multi-faceted signal analysis. Finally, we will use a multidimensional ‘quality space,’ defined by the most useful of these classifiers, as a basis for learning, modeling, and predicting the musical preferences of individual listeners, both by analyzing large numbers of actual music collections (again, mined from the web) as well as through direct subjective elicitation.

The direct product of this program will be a personalized music similarity browser. This application will be extremely useful in online music services, in-store kiosks for record stores, and for music metadata generation (Scheirer 2002). It will help to fulfill the promise of Internet music distribution and micro-publishing; electronic distribution falls short of its full potential without a method for finding interesting music to retrieve. There are hundreds of thousands of musicians who have placed their music on the Internet, yet we only hear about the few that are heavily promoted by the record industry. Facing a classic case of information overload, consumers need a way to sift through the overwhelming pile and find music to their liking. From a musician’s perspective, a good personalized music browser allows potential fans and purchasers to find their music without relying on the established record industry, which is infamously difficult to penetrate and does not always share the artist’s goals.

In a broader context, the program will have a diverse impact on fields such as machine perception, cognitive science, marketing, e-commerce, musicology, and scientific education. The techniques applied to music in this program easily transfer to machine listening in general, as well as other machine perception fields such as machine vision. Perception is the bridge between the physical world and the information world, and by building better tools for machine perception, we open the way for new devices that can record, organize and describe events that take place in the real world, including personal information appliances and security monitoring systems. The increase in understanding of human cognitive processes such as preference formation and semantic attachment will lead to increased usability in computing systems. For e-commerce, novel agents equipped with a more sophisticated understanding of personal taste can participate in bidding, shopping, and market research on the user’s behalf. In music psychology and musicology, a better understanding of how music affects people emotionally and cognitively will help explain why people listen to music. In the music business, the results of the research program could be used to build marketing tools that predict listener response to music, augmenting the data gathered from focus groups with human subjects. Finally, music recommendation and similarity browsing will be a high-visibility application of machine learning, computer science, and signal processing, which will help generate interest in the fields, particularly in young people who are the primary early adopters of digital music technology.

2 Background

In this section we discuss prior work in the field of music information retrieval, with the specific goal of motivating our approach and contrasting it with previous work. Previous retrieval systems can be roughly organized into several approaches: query-by-humming (QBH), query-by-example (QBE), and audio fingerprinting. Furthermore, approaches can be distinguished by whether they work with audio signals or at the musical score level. We also discuss prior work on recommender systems, particularly collaborative filtering.
In query-by-humming systems, the user specifies a melodic query by singing or humming into a microphone. The system performs pitch transcription on the query and retrieves items from the database with similar melodies, e.g., (Nishimura and Zhang 2002; Kosugi 2000). Generally, melody matching is done at the musical score level, that is, in a high-level representation such as standard music notation, MIDI, or a pitch-and-duration “piano roll” representation.

In query-by-example, the audio recording of a query song (or fragment) is used for retrieval. The researcher designs perceptually- and musically-motivated feature extractors, and matching is performed in the induced feature space, as in (Yang 2001) and (Wold, Blum, Keislar, and Wheaton 1996). In related work, a classification task such as artist or genre classification is performed in this feature space, such as in (Tzanetakis, Essl, and Cook 2001; Whitman, Flake, and Lawrence 2001; Berenzweig, Ellis, and Lawrence 2002a).

In audio fingerprinting, the goal is somewhat different: to find music in the database that exactly matches the audio query, in a manner invariant to transformations of the signal such as noise, audio coding, and other channel effects (Herre, Allamance, and Hellmuth 2001; Haitsma and Kalker 2002).

The earliest music retrieval systems operated at the score level, that is with transcriptions in some higher-level music notation. This work continues today, for example (Dovey 2002; Dannenberg, Thom, and Watson 1997). Generally, this work is more applicable to classical music, where transcriptions are more readily available than for popular music, because of the difficulty of the automatic transcription problem.

The transcription problem is the analog of automatic speech recognition (ASR) in the music domain; the input is audio and the desired output is the musical score. While some success has been achieved with monophonic pitch estimation systems (Goto 2001) and tempo estimation (Scheirer 1998), the more general problem of multi-pitch estimation is much more difficult (Klapuri 2001), and that does not even address the problems of identifying instrumentation, dynamics, musical structure, and embellishments.

In this project, we will take a different approach, which might be called query-by-description (QBD). Like QBE, we begin from the actual audio signal (rather than from the score level), and we induce a feature space in which to perform similarity matching. However, we then introduce an additional layer of modeling, from perceptual features to semantic or descriptive features. Early work demonstrates this approach to be quite promising (Whitman and Rifkin 2002). We will use some score-level perceptual features, but these will be transcription from audio using techniques trained with the help of MIDI alignment (section 3.2.1.

Finally, a note about recommender systems. A distinction is generally made between content-based and collaborative filtering. Collaborative filtering (Shardanand and Maes 1995) uses preference statistics from many users to recommend items that were given high ratings by users whose preference profile is similar to mine. For example, the familiar refrain on Amazon.com, “Users who bought X also bought Y.” The fundamental limitation of such systems is that they know nothing about the content of the items, and therefore cannot recommend items that have not yet been rated by many users. Our system is content-based, and thus does not suffer from this limitation.
3 Proposed Research

3.1 Overview

An overview of the project is illustrated in figure 1. The raw audio signal of individual music pieces is processed by a set of music-specific feature extractors to perform partial melodic transcription and chord-sequence recognition, and, from these, segmentation into themes and phrases. All these analyses provide input to an array of subjective attribute estimators, which map the original audio signal into a high-dimensional subjective attribute space, defined using terms and concepts associated with particular musical recordings web mining and formal subjective listening experiments. Musical preferences of a particular user are predicted via a preference mapping, a simple transformation of the attribute space that optimally conforms that space to the individual’s existing music collection.

This work brings together three main research themes, each of which is itself novel and emerging, and whose combination can far exceed comparable previous work. Firstly, we analyze the structure of the original audio signal not by relying on the short-term direct physical characteristics of the sound (such as are used in speech recognition), but rather by applying music-specific analyses to describe the music in terms of prominent melodic and harmonic content and phrase and section structure. New techniques for each of these will emphasize the kind of data-driven tools that have proven so successful in machine learning: rather than trying to describe by hand what a particular note ‘looks like’, we will train algorithms with large numbers of noisily-labelled examples.

Secondly, the stage relating the signal and its derived attributes with the descriptive, subjective terms preferred by listeners is unprecedented in this domain. Web mining will collect extensive ‘community metadata’, texts that associate particular pieces of music with sets of words. By evaluating classifiers trained for each significant descriptive term, we will find a set of robust ‘subjective dimensions’, which we can prune for optimal utility and parsimony.

Thirdly, this new ‘quality space’ will provide a malleable domain for modeling user preferences (as indicated by, say, his existing collection of recordings) as a geometric cluster. Statistical analysis of the preferences of many listeners, backed up by direct subjective experimentation, will address interesting questions about differences in taste and sensitivity to different aspects of the music. These proposed research threads are described in more detail in the following sections.

3.2 Feature calculation

This section describes the processing applied to the raw musical recording in order to generate representations that will form a solid and general basis for modeling higher-level phenomena.

3.2.1 MIDI Alignment for Automatic Transcription of Music

Amateur listeners have difficulty transcribing real-world polyphonic music, replete with complex layers of instrumentation, vocals, drums and studio effects. Indeed it takes an expert musician with a trained ear and a deep knowledge of music theory to be able to accomplish such a task. Thus, it
Figure 1: Overview of the music preference modeling system.

is natural that the transcription problem be posted in a statistical pattern recognition framework, as in (Walmsley, Godsill, and Rayner 1999).

A fundamental obstacle in both the training and evaluation of transcription algorithms is the lack of labeled ground-truth data from real-world audio. Some music is available as MIDI, a digital representation that encodes every note event. MIDI files can be fed to synthesizer programs to generate an audio signal that is a fair musical approximation of the original piece — although the ‘synthetic’ instrument tones are easily distinguished from the original recordings. Comparing the output of the synthesizer with the MIDI events gives a very precise ground truth of notes in the audio stream; however, the simplified audio signal is in general not a good match for training recognizers able to recognize note events in real music recordings from CD. Therefore current research is limited to unsupervised learning methods, or systems in which the feature structure associated with particular notes is more or less explicitly provided by the system architect, and quantitative results (i.e. a Note Error Rate) can generally only be provided for toy problems.

Due to the well-known harmonic structure of musical tones, providing explicit models is relatively feasible, and recent systems have shown considerable promise with this approach (Dannenberg and Hu 2002; Klapuri 2001; Goto 2001). However, there may be subtle interactions between particular feature representations and the manifestations of each note, and the interaction between notes is either ignored or only weakly addressed.

In order to overcome this difficulty, we propose building a corpus of labeled real musical recordings by aligning songs with their MIDI transcription. MIDI files are another kind of community metadata, transcribed by professionals for Karaoke machines, amateur musicians paying deference to beloved artists and music students honing their craft in the same vein as an apprentice painter replicating the work of Picasso. We have spidered online MIDI search engines to find transcriptions to match many of the songs in our local music repository.

In order to align a song with its transcription we create a similarity matrix, with each point \((i, j)\) populated with the inverse cosine distance between time-frequency analysis features of the \(i\)th window of the original performance and the \(j\)th window of the synthesized MIDI. The features are chosen to highlight pitch and beat/note onset times, and to normalize away slowly-varying systematic difference between the two signals, by differentiating along both time and frequency axes. Next, we use dynamic programming to find the lowest-cost path between the starts and ends
of the sequences along the similarity matrix. Finally, this path is used as a map to warp the timing of the MIDI transcription (which will respect the general rhythm of the original, but not its precise timing) to match that of the actual CD recording.

The result of this processing is a substantial collection of real, well-known pop music recordings with almost complete note-level transcriptions aligned to within a few milliseconds of accuracy. This can be used as training data for a robust classifier, without any explicit (and suspect) signal models. For instance, by collecting together all the audio corresponding to note C4, across all the different instruments and against many different backgrounds, we can train a standard statistical pattern recognizer to identify for itself the common and reliable features of this note, which may depend on our feature representation in subtle ways not reflected in explicit models.

Figure 2 shows an example of the alignment between MIDI and CD renditions of “Africa” by the band Toto.

3.2.2 Chord transcription

Transcribing music into individual notes is difficult because an unknown number will be present simultaneously; our best models for classifying signals, such as speech recognition, work by finding a single class label for each segment. If we could find a useful labelling of the music signal in terms of a succession of ‘global’ labels, a range of sophisticated and mature tools could be applied.

One such label sequence is the chord progression. Conventional western music can usually be characterized as a sequence of chords, each lasting between a beat and several bars, but only one at a time. Simple musical transcriptions may consist of the words, the melody (i.e. the pitches sung by the singer) and a sequence of chord codes such a G or Amin7, which a trained pianist or guitar player can immediately reproduce. Although many songs will share the same chord
sequence, they would have an immediately obvious ‘similarity’ if played alongside one another. “Let it be” by the Beatles and “No woman no cry” by Bob Marley is one pair of well-known songs with nearly identical chord sequences. Thus, automatic extraction of the chord sequences forms a useful foundation for automatic music similarity and preference prediction.

Prior work in this area has generally decomposed the problem as one of identifying individual notes (i.e. full polyphonic transcription), then identification of chords from the note sets (not a trivial problem even when the notes are perfectly known) (Kashino, Nakadai, Kinoshita, and Tanaka 1998; Raphael 2002; Pardo and Birmingham 2001). The limited success in this approach can be ascribed to the ‘double jeopardy’ of having to accommodate errors in the first level transcription when making the second level chord assignment; also, if the goal is simply to recover the chord identity, recovering the actual notes making up that chord is much more detail than is required. Many different combinations of notes will carry the same chord, and it would be more parsimonious to recognize that chord on the basis of its common characteristics, rather than going through an intermediate note representation.

We are faced with the problem of building statistical models of the acoustic properties of individual chords, starting from a set of acoustic examples along with their chord sequences (e.g. from play-along transcriptions), but where the precise time alignment of the chords to the sound waveforms is not known - only the sequence in which they will occur. This is an exact analog of the common problem in speech recognition, where the bulk of the training data may be transcribed into word sequences, but timing information is not available within each utterance segment. The Baum-Welch EM training algorithm used in speech can be applied equally well here, to simultaneously learn the feature distributions for each chord label, and the segmentation of the training examples into those chords.

In preliminary investigations, we took a small corpus of 20 songs from two Beatles albums, and obtained chord sequences from a large archive of chord transcriptions available on the internet (http://www.olga.net/). Using the HTK speech toolkit, we trained models for 7 major chord classes (for 12 possible keys) using 18 of the tracks. The features were constructed to make the signature of the same note played at different octaves appear similar as in (Fujishima 1999) and (Bartsch and Wakefield 2001). Testing on the remaining two achieved an (alignment) frame error rate of 16.7% when recognition was constrained to the correct chord sequence, but increasing to 76.6% on unconstrained recognition. These results confirm the viability of the approach and our feature set (standard MFCC features achieved no lower than 73% frame errors even when constrained to the correct chord sequence) at the same time as underlining the need for a larger training set in order to generalize. (It may also be that the music of the Beatles, which is in fact quite harmonically sophisticated, may not be the simplest starting place).

3.2.3 Phrase segmentation

It is not so much the individual, rapidly-changing note and chord identities that lead to the perceived nature of the music, but rather the larger patterns (melodies and progressions) formed by their sequences. To describe acoustic recordings at this level, we will need to segment recordings into musical phrases, the next highest level of temporal structure.

Segmenting a continuous signal into sections that ‘belong together’ is a common problem,
directly analogous to segmenting recorded discussions as the speaker changes (an important task in speech recognition). In (Chen and Gopalakrishnan 1998), the Bayesian Information Criterion (BIC) is used to decide whether the likelihood gain achieved by dividing a single stretch of data into two subsequences either side of a boundary is adequate to justify the doubling of the number of parameters involved in modeling two segments instead of one; we have recently used the same principle at a higher scale to segment recorded meetings into ‘episodes’ that involve particular subsets of the participants (Renals and Ellis 2003). With an appropriate probabilistic model for the underlying spectral, melodic and/or harmonic features, BIC can also be used to segment music into sections that have distinctly different properties.

Other approaches to music segmentation include using self-similarity matrices of the same basic form as 2, but comparing a signal to itself (Foote 1999). Segment boundaries can be placed at locations of minimum self-similarity along the leading diagonal. This display also reveals repeated segments within a piece as off-diagonal structure, leading to the identification of ‘chorus’ and ‘verse’ (Bartsch and Wakefield 2001). Repeated segments can also be identified by direct clustering of small windows of the signal to see which pieces occur most frequently (Logan and Chu 2000).

Melodies and harmonic sequences extracted from each segment can be matched, exactly or approximately, between pieces to find popular forms or direct quotations. As a basis for music similarity modeling, we will seek some kind of data reduction (such as projection onto a set of linearly-combined ‘eigenmelodies’ that represent an underlying theme with increasingly fine detail. This way, arbitrary sequences can be reduced to a low-dimensional, well-behaved feature space.

### 3.3 Subjective Qualities of Music

Our analysis of the audio signal will provide several representations that correspond more closely to features of importance to listeners, but they are still essentially describing the manifest attributes of the music; at a level above this, we have more general subjective qualities, which are closer to the common properties shared by the music liked or disliked by a given individual.

#### 3.3.1 Defining and grounding musical concepts

When we speak of ‘music intelligence’ and ‘music understanding,’ we imply that there is some meaning to music that we wish to capture in a computational framework. To enable an intelligent device to recommend music to a user, the meaning of both the set of signals to sort through and the nature of the query is needed. But how can meaning be represented to a computer?

In current music retrieval and understanding systems, labels are hand-selected and applied to some signal-derived feature space. For example, in the case of a genre detection system such as the one described in (Tzanetakis, Essl, and Cook 2001), the authors ‘learn’ a relation between acoustic features and their constituent genre tag using various machine learning classifiers. Success is measured by predicting genres on a new held-out test set of audio features. Other systems that operate in this highly supervised model claim successes in such tasks as artist identification (Whitman, Flake, and Lawrence 2001), style recognition (Dannenberg, Thom, and Watson 1997)
and nationality detection (Chai and Vercoe 2001).

While this model has certainly proved successful on small, domain specific and known databases, we consider a more naturalistic model of music intelligence that emulates the complex human model of describing, discovering and classifying music. In the case of genre detection, the hand labels are taken from a marketing-driven hierarchy of musical style. Most record stores segment their selections into “Rock/Pop,” “Classical,” “Rap and R&B,” and “Jazz.” By training a regression machine to segment music into these labels, we believe you are only forcing a feature space to conform to some non-musical sale-based structure. Many of these labels have no acoustic underpinning at all—rather, they reflect certain cultural features. Even if we could create a multi-modal classifier that understood both, as in (Whitman and Smaragdis 2002) – we can only claim that we can teach a computer to predict marketing delineations.

Rather, we propose that music intelligence be evolved from the ground up. Human music preference has little correlation to notions of genre or style; those are often a posteriori observations used to cluster music for convenience in browsing. But if a system could investigate the complete and complex feature space of music—both cultural and acoustic— it could infer its own styles, draw connections currently missed by the record industry, and allow for a far better music search and retrieval experience.

Our proposed model is influenced by studies in computational semantics and linguistics. Borrowing from the “symbol grounding problem” proposed by (Harnad 1990), we aim to attack the situation that no music understanding system yet actually understands music. Harnad’s main complaint is illustrated in his “Chinese dictionary” analogy in which you are asked to understand a new language given only a dictionary written in and for that language. While you may know that some symbol is defined by a set of other more primitive symbols, you cannot function without a relation between a symbol and the outside world. Relating a symbol to the outside world is termed ‘grounding,’ and computationally, one can ground terms to perceptual input via any sort of machine learning apparatus. Work has been done to allow systems to understand color and movement (Roy 1999) or objects in a static image (Duygulu, Barnard, Freitas, and Forsyth 2002; Barnard and Forsyth 2000). In the case of supervised genre detection as described above, we can link the symbol ‘electronic’ to a set of perceptual symbols by hand-labelling a set of music as electronic. But this connection falls apart if studied robustly: can this system suggest something similar or opposite to ‘electronic?’ Can it provide canonical examples of or define ‘electronic?’ Can it realize that ‘electronic’ is at best a loose grouping of culturally-defined features invented purely for increasing record sales?

A more evolutionary approach to grounding music would involve a completely unsupervised mechanism for learning the meaning of music and inferring new relations. Our model starts by assuming a blank slate of perception and knowledge. The aim is to slowly grow understanding by first learning the descriptions of varying musical input. In (Whitman and Rifkin 2002) we make a first pass at this problem by linking music to autonomously collected textual descriptions of music. In this system, the computer both ‘listens’ to the music and concurrently ‘reads’ about the artists. As a result it learns the sound of various musical descriptors, as in Figure 3. This more organic approach to learning allows for an enormous amount of expressivity in understanding: instead of being limited to a vocabulary of five to ten genre terms, our system can describe music by using any combination of 100,000 terms collected by our ‘reading’ agents. (This representation of internet-
wide musical description, which we call “community metadata,” is described in (Whitman and Lawrence 2002).)

Using this method, we can claiming better understanding of musical signals. Any number of music retrieval systems could benefit from these semantically attached terms due to their expressive power. The most obvious application is a ‘query-by-description’ system, where users can search for music by merely asking for it. Once our system begins to ground terms in music, it can start automatically labelling new audio that has yet to hit the public’s ear, in effect an automatic reviewer that can be tuned to a specific user model.

Future work on music grounding is underway; we are currently studying how to automatically learn parameter spaces in perception (such as loud to soft, or fast to slow.) (Whitman 2002) These ‘feature knobs’ can be integrated into retrieval interfaces. We are also studying in depth the class of lexical terms that are semantically attached (Whitman 2003)– for example, a text retrieval system that needs to return topical documents on music would be better off limiting itself to grounded terms like ‘acoustic’ and ‘female,’ using grounding as a noise reduction technique.

### 3.3.2 Musical attribute estimators

This section describes our method for mapping perceptual space into a semantic attribute space and how to measure similarity between musical objects in that space.

The basic approach is to train a set of classifiers, each of which is tuned to recognize membership in a musically-relevant semantic category. If we collect the output (posterior probabilities) from several such classifiers, the result is a new vector of features, perhaps of lower dimension, where each dimension represents soft membership in one of the attribute classes. In other words, points in attribute space are vectors of posterior probabilities of membership in the attribute classes, given the input:

\[
(p(\omega_1|x), \ldots, p(\omega_M|x)),
\]

where \(\omega_i\) represents the \(i^{th}\) attribute class.

From a machine learning perspective, the attribute classifiers can be seen as nonlinear feature extractors, where the nonlinear function is obtained by machine learning techniques. Seen this way, this work is related to the work of Bollacker and Ghosh (Bollacker and Ghosh 1998) on “supra-classifiers” and knowledge reuse. Another related technique is tandem acoustic mod-
eling for speech recognition, where the output from neural networks trained to recognize phone classes are further modeled using GMMs (Ellis and Gomez 2001) (Hermansky, Ellis, and Sharma 2000) (Sharma, Ellis, Kajarekar, Jain, and Hermansky 2000). Slaney (Slaney 2002) uses a similar technique for content-based audio retrieval.

We have implemented a small system using 10-20 attribute models, trained with genre and artist labels. For classifiers, we used neural networks (multi-layer perceptrons), and for input, we capture the spectral shape of the audio with mel-frequency cepstral coefficients (MFCCs) computed over a short time window.

The resulting attribute space was used to achieve 38% accuracy on a 400-class artist classification task (Berenzweig, Ellis, and Lawrence 2002b), considerably more difficult than the 21-artist set which was the largest previously reported (Berenzweig, Ellis, and Lawrence 2002a). (Although a higher accuracy of 65% was reported in this earlier work, it would not have scaled up to a larger set.) In an evaluation based on human subjects’ judgments of music similarity (described in section 3.5.1), a centroid-based similarity measure in this space (also described below) outperformed several similarity measures derived from sources of human opinion, such as collaborative filtering and webtext, and performed comparably to a measure derived from expert opinion.

In this preliminary work, we trained the classifiers using hand-picked genre labels. For the full project, we will apply the same approach, but using semantic labels gathered from the Internet as described in the previous section. By generating a very large number of candidate dimensions and labels, we can choose among them for subjective quality attributes that are both accurately predicted from acoustic features (indicating they really do describe something about the signal), at the same time as optimizing the coverage and independence of the attributes, so the most complex musical space can be constructed with only a modest number of dimensions.

3.3.3 Similarity Measures: Comparing Clouds

Because feature vectors are computed from short time segments of audio, an entire song induces a cloud of points in feature space. The cloud can be thought of as samples from a distribution that characterizes the song, and we can attempt to model that distribution using standard statistical techniques. Extending this idea, we can conceive of a distribution in feature space that characterizes an entire album, an artist’s repertoire, a genre, or any other set of music. Modeling and comparing these distributions is the subject of this section.

Gaussian Mixture Models (GMMs) are a good choice because of their ability to model complex multi-modal distributions and the ease with which they are trained using the Expectation-Maximization (EM) algorithm (Bilmes 1998).

To measure similarity, we would like to use the Kullback-Leibler divergence, which is the natural way to define distance between probability distributions (Kullback 1968). However, we run into difficulty because the KL-divergence between mixture densities does not have a closed form in general, and in particular for GMMs no closed form is known (Vasconcelos 2001).

In preliminary work, we examined several candidate measures of similarity: two approximations to the KL-divergence, and two simpler methods where we first reduce clouds to a single point and then take the Euclidean distance.

In one approach, the KL-divergence between all pairs of mixture components is calculated.
The closed form for the KL-divergence between two single Gaussians \( p(x|\theta_1) = \mathcal{N}(x|\mu_1, \Sigma_1) \) and \( q(x|\theta_2) = \mathcal{N}(x|\mu_2, \Sigma_2) \) is well known to be

\[
D(p||q) = \frac{1}{2} \left[ \log \frac{\Sigma_2}{\Sigma_1} + \text{tr} \left( \Sigma_1^{-1} - \Sigma_2^{-1} \right) + \text{tr} \left( \Sigma_2^{-1}(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T \right) \right]
\]

To compare two mixtures \( P \) and \( Q \) with components \( (p_1, \ldots, p_K) \) and \( (q_1, \ldots, q_K) \), we first calculate the KL-divergence between all pairs of components, \( D(p_i||q_j) \). Now we take the mean of the smallest divergence for each component of mixture \( P \):

\[
\hat{d}(i) = \min_j D(p_i||q_j) \quad \quad \quad D_{\text{meanmin}}(P, Q) = \frac{1}{K} \sum_{k=1}^{K} \hat{d}(k)
\]

Note that this method does not use the mixture coefficients, \( \pi_k \), and thus is an extremely crude stand-in for the true KL-divergence. Further development of this method is required, perhaps connecting it to the earth-mover’s distance (REF?).

In the second approach, we sample \( L \) points from distribution \( P \) and then compute the likelihood of model \( Q \) given the samples. The (log) likelihood of a GMM with parameters \( \theta \), given the observed set of samples \( \mathcal{X} = (x_1, \ldots, x_T) \) is

\[
l(\theta|\mathcal{X}) = \sum_{t=1}^{T} \log p(x_t|\theta) = \sum_{t=1}^{T} \log \sum_{k=1}^{K} \pi_{ik} \mathcal{N}(x_t|\mu_{ik}, \Sigma_{ik})
\]

where \( \theta = (\pi_1, \ldots, \pi_K, \mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K) \) are respectively the priors, means, and covariance matrices of the \( K \) mixture components for the mixture model, and \( \mathcal{N}(x|\mu, \Sigma) \) is the Normal distribution.

Sampling is theoretically cleaner, and leads to a nice interpretation of the similarity measure as the likelihood of confusing points generated by model \( P \) as having been generated by model \( Q \). However, in high-dimensions, a large number of samples required to adequately characterize the distribution is large, and early experiments were not successful. However, more work is required to establish or discount the viability of this approach.

We also examined two simplified similarity measures (more precisely, distance measures) in which we first reduce clouds to a single point and then use a point metric. The first of these, the centroid distance, is the Euclidean distance between the centroids of two objects in attribute space. The second point-measure uses a histogram of the most active attributes over time. The histogram of the most-active attribute model at each time frame is normalized into an activation score which measures the deviation of the model’s activity on object \( P \) from its mean activity over all objects.

### 3.4 Musical preference modeling

Having defined attribute space and how to map perceptual features into semantic categories, we turn to the original motivation: modeling musical preference. As mentioned previously, our approach is to define a personalized similarity space for each user. For information retrieval, we can
use the personalized space for indexing and responding to queries. To make recommendations, we then choose music that is close to the user’s collection in this space. (After all, recommendation is just information retrieval where the query is an entire collection rather than a single song.)

The first step is to cluster the user’s collection in attribute space. Perhaps the user has heterogeneous interests, and it may not be appropriate to analyze the entire collection on the whole. For instance, Adam likes certain types of hip-hop and certain female singer-songwriters, but we don’t believe that his reasons for liking those hip-hop artists and not others would explain his taste in female songwriters.

On each cluster, the next step is to use principal component analysis (PCA) to find the linear transformation (a rotation and a squish) that best captures the variance of the cluster. However, we are in fact looking for components of the least variance, because they best “explain” the cluster. What we mean is, if we were doing classification, the components of least variance in the cluster would be most helpful in separating the cluster from the rest of the space. PCA has ordered the columns of the transformation matrix by variance, and we are interested in only the \( N \) components of least variance. We can zero out the columns corresponding to the other components.

Remember that the goal is to find a transformation into a personalized attribute space where Euclidean distance matches the user’s notion of similarity, as expressed indirectly through their choice of music. Therefore, we may want to weight the selected components in such a way that compresses the more important dimensions and expands the less important ones, for example, by weighting each component proportionally to the cluster variance along it.

After having derived several transformations by analyzing each cluster, then we need to combine components from different clusters to produce a single personalized similarity space. The columns of each transformation matrix represent a linear combination of attribute-space dimensions that produces a single component dimension of the new space. We have already selected the columns of interest and zeroed-out the others. To combine matrices, simply concatenate the non-zero columns from each matrix. It may be beneficial to weight the components from each cluster by that cluster’s representation in the collection. Note that components from separate clusters may end up being highly correlated, so some redundancy checking will probably be a good idea. If highly correlated components are found, they can be merged, combining their weights by averaging or summing.

### 3.5 Evaluation

Evaluation is difficult in this kind of domain because there is no clear ground truth for music similarity (Ellis, Whitman, Berenzweig, and Lawrence 2002). We are obliged to use objective measures whose relevance is questionable, and/or to conduct subjective judgment and user tests.

Typically, music recommendation systems are evaluated using leave-one-out methods: a portion of the user’s collection is withheld from training, and the investigator examines the number and rank of recommended songs that match the withheld set. This type of evaluation is useful but problematic. It is useful because it is quantitative, straightforward to interpret, and simple to perform if enough data is available. User collections are easy to obtain by mining peer-to-peer filesharing networks. However, no explicit rating information is available; the typical simplifying assumption is that the user likes every song in her collection equally, yet the collections have rarely
been constructed with the level of care that suggests.

3.5.1 Evaluating Similarity Measures

A more direct method is to evaluate candidate similarity measures by comparing them with data gathered from human subjects in a survey about artist similarity. In (Ellis, Whitman, Berenzweig, and Lawrence 2002), we presented subjects a list of 10 artists \(a_1, \ldots, a_{10}\), and a single target artist \(a_t\), and asked “Which of these artists is most similar to the target artist?” We interpreted each response to mean that the chosen artist \(a_c\) is more similar to the target artist \(a_t\) than any of the other artists in the list \(a_1, \ldots, a_{10}\), if the artists are known to the subject. By running this survey as an anonymous web page, we were able to collect more than 20,000 judgments in a few weeks, apparently because people found participating in a survey about popular music quite enjoyable. (We have made this data available for anyone else in the community to use, via http://musicseer.com/results/).

For the more detailed questions of preference and similarity not among artists but among individual tracks, we will likely have to employ more controlled tests. We are investigating how to make inferences based on users’ listening patterns from our prototype similarity-based browser, stocked with thousands of noncommercial recordings (Berenzweig, Ellis, and Lawrence 2003). Depending on the most pressing questions and feedback that emerge from prototype browsing systems, we will incline towards a tiered subjective data collection strategy, with large amounts of relatively indirect data being collected as a ‘side-effect’ of a web music browser that we hope will be used for its own sake, supplemented with much more specific experimental results from controlled psychoacoustical tests conducted in the lab.

4 Team and Plan

This work will be centered in Columbia University’s Electrical Engineering department, at the Laboratory for Recognition and Organization of Speech and Audio (LabROSA) which was established by PI Ellis in 2000. Two graduate students will work there full time, one on extracting features from the music signal, and one on similarity modeling and browsing. In addition, a third graduate student from the MIT Media Lab (Brian Whitman) will also be working directly on this project, although he will be funded separately through the Media Lab (see attached letter from his advisor, Prof. Barry Vercoe). Finally, Dr. Beth Logan of HP Labs in Cambridge has been working on issues of music similarity and browsing for some time and has agreed to collaborate with us very closely on this project, dedicating up to 50% of her time to this work (see her attached letter). The involvement of HP as an industrial partner with a direct commercial interest in music access technology will provide a practical and pragmatic influence on the project, as well as alternative sources of subjective evaluation results.

Our situation at Columbia confers a number of specific advantages: Eben Moglen of the Columbia Law School is a leading authority on issues of intellectual property and copyright implications arising from new media; he has been generous in giving us informal advice in the past. We can also draw on the strength of the local Psychology faculty to guide and advise us in the development of subjective tests (see the attached letter of support from prominent psychoacoustics authority Prof. Robert Remez). Finally, Columbia has a strong and innovative Computer Music
Center with which we have strong ties, and which itself is deeply involved in the vibrant musical culture of New York City.

5 Broader impact

Music holds a powerful attraction for all people: This proposal grew out of several spontaneous, unsupported efforts begun as student projects for DSP and audio processing classes. Ongoing research in musical signal analysis will providing new results and materials for this class, and we are already discussing the establishment of a new class devoted entirely to engineering and information processing aspects of music.

Exciting and accessible results in music browsing will afford this project a high profile in a much broader community. LabROSA is a regular fixture for the Engineering Open House events organized by the school for prospective students: sound and music has an immediacy for teenagers who are considering engineering, and this area communicates engineering’s allure and relevance. LabROSA is in the process of establishing a relationship with a local high school for a small-scale mentoring program.

We have found that public interest in this work can be intense. Our pilot project in collecting artist similarity ratings collected two orders of magnitude more data than we expected, largely due to a number of stories that appeared in the ‘webslogs’ of chance observers, whose curiosity amplified by the many people who subsequently participated in the survey. This and other media interest will help the project to serve an ambassadorial role representing the aesthetic and sensitive side of technical research.

6 Conclusion

Music preference is a deep and complex behavior, which may explain why there has been no previous effort to quantify the way listeners use musical structure to form their tastes. However, it is only by investigating this question systematically that we can gain a clearer understanding of how much and how accurately personal preference can be explained by quantitative models.

By fusing audio structure analysis, community metadata, and geometric preference models, we hope to explain significant components of musical preference, at the same time developing new music search and browsing tools unlocking a huge reserve of unmarketed music for casual listeners. The goal of building a functional model of such an involved and subjective phenomenon will establish a paradigm that can be reused in the large number of analogous problems that occur everyday, from movie reviewing to designing the look of new cars.

7 Results from prior support

PI Ellis is involved a co-PI on one current NSF project: NSF IIS-0121396 ($1,402,851), Title: ITR/PE+SY: Mapping Meetings: Language Technology to make Sense of Human Interaction, Award period: 2001-09-01 to 2005-08-31, PI: Nelson Morgan, International Computer Science
Institute. This project is concerned with the application of speech recognition and other automatic signal analysis techniques to extracting information from recordings of natural, unconstrained meetings between human participants. The one graduate student in LabROSA supported on this grant is currently looking at unsupervised clustering of the extensive data so far collected as a way to define and locate “interesting” events in the recordings, and there will likely be transfers of these techniques into the current project. Several publications arising from this work are currently in review; the initial data collection and general goals are described in (Morgan, Baron, Edwards, Ellis, Gelbart, Janin, Pfau, Shriberg, and Stolcke 2001).
References


Whitman, B. (2002). Grounding gradations from perception and a professional. *In process*.


Whitman, B. and S. Lawrence (2002). Inferring descriptions and similarity for music from community metadata. submitted.


