Modeling Music Similarity: Signal-based Models of Subjective Preference  
Daniel P.W. Ellis, Electrical Engineering, Columbia University

Summary

Music preference is highly subjective and individual, yet it is a very powerful experience on the part of each listener. This project will investigate the hypothesis that music preference can be predicted by acoustic-based similarity between a novel piece and representative examples from the listener’s existing collection. The major challenge in this investigation lies in finding the appropriate representations for the music signal and the comparison techniques that predict subjective similarity most closely.

The ideal application arising from this project would be a system that takes information describing the particular musical taste of a listener, such as their music collection or logs of their personal listening habits, and uses this to construct an abstract description of their musical preferences. This description can then be matched against similar representations derived from an unlimited amount of new music, to provide personalized recommendations and navigation without any intermediation through marketing categories or other limitations. Critically, all the information is derived only from the acoustic signal: unlike collaborative filtering (which relies on matching individuals against others who share their interests), the system would be equally applicable to well-known music and to completely obscure artists who happen to make their recordings available.

This work will make extensive use of machine learning techniques, to infer models and parameters for incompletely-defined attributes from collected examples. This applies both at the highest level, where a listener’s musical taste is modeled as a distribution in an appropriate “subjective musical feature space”, and at various preceding levels, such as phrase segmentation and chord transcription, which will be learned from labels originating (in some form) from human informants.

**Intellectual merit:** This project will develop a set of automatic analysis tools for segmenting music signals into self-consistent phrases, and extracting spectral, rhythmic, harmonic, and melodic ‘signatures’ from each of these phrases. Each of these attributes will be mapped into a high-dimensional perceptually-relevant space for comparison and clustering. Analysis of people’s listening patterns will provide a functional definition of preference, and this ground truth will allow us to develop signal-based music preference predictions. All of this work is either completely novel, or is only at a rudimentary level at present.

**Broader impact:** The project will support a nascent cross-disciplinary collaboration with the music department for classes to involve a wide group of students in aesthetic-engineering projects. There will also be a workshop organized toward the end of the project to bring together interested researchers from the US, Europe and Japan. The motivating vision of this project, of an automatic system for efficiently connecting music makers with listeners based purely on the sound of the music, has the potential to revolutionize how every consumer of popular music finds new material.
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 exclusively from the recorded signal.

This project addresses the scientific question of whether an individual’s preference for music can be usefully predicted based only on the similarity between the signal content of the new music in comparison to a known collection of music ‘liked’ by the listener. This immediately leads to the question of how to measure this similarity, and a host of problems regarding the appropriate analysis and representation of the signal, how to collect and infer ground-truth concerning individual music preferences, and how to evaluate the resulting models. It is these questions that the research proposed below seeks to answer.

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 given that 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 significantly (although not exclusively) 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 to consider a wide range of signal-to-similarity transformations, some based on attempts to extract known listener-relevant attributes such as notes and rhythm, others more blindly trained from subjectively-labeled examples to parameterize general purpose transformation systems (such as artificial neural networks) to best approximate the subjective judgments.

This project will make several key contributions to the newly-emerging and active area of music content analysis. Firstly, we will advance current work in structural analysis of musical signals (i.e. transcription of themes, harmonies and rhythms) using data-driven machine-learning techniques that are unprecedented for these kinds of signals. Features derived from these anal-
yses at the phrase level will be mapped into subjectively-relevant ‘anchor spaces’ through sets of standard classifiers that gauge the match of a particular fragment to small exemplar sets characterizing human-defined categories; the pattern of likelihoods returned by this set of classifiers form points or clusters in a new space that shares the basic geometry of subjective assessments – somewhat independent of the exemplar categories provided they span an adequate range. We will use this a multidimensional ‘quality space’ as a basis for learning, modeling, and predicting the musical preferences of individual listeners by processing actual personal music collections (obtained from volunteers and the web), and by analyzing the behavior described in listening logs from personal digital music devices (again, contributed by volunteers). In addition to providing automatic methods for predicting preference from signals, this analysis will also give us a much deeper understanding of the nature and variety of music preference experience.

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 made their music freely available on the Internet, yet we only know about the few that are promoted by the record industry. Facing a classic case of information overload, consumers need a way to sift through this 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 computer vision. Perception is the bridge between the physical world and “information”, 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 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 and corroborating data gathered from focus groups. 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

2.1 Music content analysis

When we listen to music, our impressions of its content (in terms of instruments, notes, phrases etc.) are so strong that it seems as though automatic extraction of these attributes might be quite easy. A long history of small achievements won with great efforts belie this impression: the music
signal, with its coincident harmonics and synchronized events, is an extremely challenging analysis problem.

Starting with (Moorer 1975), systems for pitch transcription (i.e. identifying the overlapping note sequences) have fascinated successive generations of researchers, leading to recent systems that usefully capture one or two constrained melodic streams (Goto 2001), and show promise transcribing all voices (Klapuri 2001). These systems typically search for an explanation of the observed signal in terms of idealized pitch models, either in the frequency domain, or even directly in the time domain (Walmsley, Godsill, and Rayner 1999). Efforts to extract further information, such as fine timing details, instrumentation, or embellishments, await more robust solutions to the basic pitch detection problem.

Rhythmic transcription has been studied as a separate issue, starting with the basic problem of finding the fundamental pulse or tempo (Scheirer 1998), and extending into issues of identifying the patterns of different drum sounds (e.g. bass and snare (Goto and Muraoka 1995)). Extraction and modeling of rhythmic ‘feel’ in acoustic recordings has received attention (Bilmes 1992; Laroche 2001), but this information has not been used as a basis for broader music similarity or classification.

Harmonic transcription (i.e. converting audio into a chord sequence) has been proposed as a more natural and perhaps more tractable alternative to note transcription, but has still proved very difficult except for highly-controlled acoustics (such as a known electronic keyboard) (Fujishima 1999). The complexity of musical harmonic theory makes it nontrivial to derive the correct harmonic labels even when the notes are exactly known, let alone when starting from an acoustic signal (Raphael and Stoddard 2003).

Phrase segmentation of acoustic recordings has mainly been studied for the purposes of extracting highly salient snippets – “audio thumbnails” – to summarize a piece of music. The idea that a chorus segment would be repeated most frequently within a piece underlies several systems (Logan and Chu 2000; Bartsch and Wakefield 2001). An algorithm for an exhaustive decomposition into all repeated units is presented in (Chai 2003).

The research we propose seeks to develop all these attributes as potential factors in subjective similarity and preference.

2.2 Music similarity

In terms of comparing different audio versions of music, the most thoroughly studied area is audio fingerprinting – that is, identifying a short, possibly corrupt, fragment as originating from a known existing recording. This problem has proved surprisingly tractable (Herre, Allamance, and Hellmuth 2001), and commercial services exist in the UK and elsewhere that will “name that tune” over a telephone held up in a noisy bar when desired music is playing in the background (Wang 2003).

“Softer” forms of music similarity, such as identifying works of the same artist or genre, or different performances of the same piece, prove much harder. The general idea, common to our approach, is to define some perceptually- and musically-motivated feature space in which to perform matching e.g. (Wold, Blum, Keislar, and Wheaton 1996; Tzanetakis, Essl, and Cook 2001; Whitman, Flake, and Lawrence 2001; Logan and Salomon 2001; Pampalk, Rauber, and Merkl 2002; Berenzweig, Ellis, and Lawrence 2003). In general, these systems use relatively simple signal features (often derived from speech recognition) and take a musically-impoverished view of each recording as a distribution of the features derived from fixed short time frames.

A major obstacle in this work is evaluation, since satisfactory “ground truth” for subjective sim-
Figure 1: Overview of the mapping from musical signal to subjective anchor space. (Shaded numbers refer to the proposal sections describing each stage.)

Clarity is hard to define. We have studied this problem (Ellis, Whitman, Berenzweig, and Lawrence 2002; Berenzweig, Logan, Ellis, and Whitman 2003) and will return to it below.

2.3 Music information retrieval

Music information retrieval is a new meeting-ground for engineers, musicologists and librarians; the annual International Symposium on Music Information Retrieval (ISMIR) started in 2000 and drew more than 130 participants in Baltimore last year. The most popular paradigm is query-by-humming systems, in which 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., (Kosugi 2000; Birmingham, Dannenberg, Wakefield, Bartsch, Bykowski, Mazzoni, Meek, Mellody, and Rand 2001; Pauws 2002). 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, and the problem of converting actual musical recordings into this form is often sidestepped by working with ready-transcribed versions (such as performances captured from electronic keyboards).

2.4 Music recommendation

Existing work in music recommendation is mainly limited to “collaborative filtering” systems (Shardanand and Maes 1995) which use preference statistics from many users to recommend items that were given high ratings by users with similar profiles, epitomized by the Amazon.com refrain, “Users who bought X also bought Y.” The fundamental limitation of such systems is that they do not make any use of signal content, and cannot recommend items that have not yet been rated by many users. This is a particularly severe limitation for the kind of scenario we are considering, where we wish to make recommendations from among the vast sea of obscure music on offer.

3 Proposed Research

An overview of the project is illustrated in figures 1 and 2. The audio signal of individual music pieces is processed by a set of music-specific feature extractors including partial melodic transcription, chord-sequence recognition, and beat-pattern extraction. The recording is then segmented
Figure 2: Using the anchor space mappings to model user taste and predict preference of new music.

into self-consistent fragments, based on these features, a self-similarity analysis, and cues trained from manually-segmented examples. The music is now represented as a set of shorter segments with nominally consistent properties.

These variable-sized signal-based feature sets are then mapped into a fixed-size subjective ‘quality space’ through a set of anchor model classifiers. Each anchor model estimates the likelihood that the phrase segment belongs to a particular class for which it has been trained from a small set of examples; these classes might be genres, particular artists, instruments, etc. The net result is a point in feature space whose dimension is fixed at the number of anchor classifiers employed, and in which closeness corresponds to similarity along each of the listener-relevant dimensions defined by the anchors.

Figure 1 occurs twice as a module in figure 2, which illustrates how the musical preferences of a particular user are modeled by analyzing their listening habits to identify the kinds of music they most like (and, potentially, finer gradations and dependencies of musical preference). These examples then define a distribution in the subjective feature space; similarity between these distributions and the analysis of novel musical pieces predicts the user’s preference for the new music.

Many aspects of this work are unprecedented in music similarity modeling, including: High-level musical features relating to pitch, rhythm, and chords; segmentation of the music into phrase-fragments, rather than comparing undifferentiated wholes; using “anchor space” models to map the segments into a perceptually-conformal space; ground-truth subjective preference data extracted from personal collections and listening logs; individual preference models obtained by warping the anchor space to suit each listener; and preference predictions and similarity browsing used to discover new, little-known music. Our extensive use of machine learning to infer classification and structure from training data in preference to ‘hand-defined’ rules, particularly in the area of extracting musical features, is another key contribution.

### 3.1 Feature calculation

The foundation of the similarity comparison lies in the musically-relevant information we can glean from the acoustic signal. We will employ a broad set of features ranging from close relatives of the signal through to more abstract, music-based attributes.

#### 3.1.1 Spectro-temporal features

The raw spectro-temporal properties of the music, roughly corresponding to the information present in early stages of the auditory system, are the foundation of all analysis, and certain attributes such
as the nature of the instrumentation may be successfully derived from them with little further analysis. Current work in music similarity uses such features exclusively (Tzanetakis, Essl, and Cook 2001; Whitman, Flake, and Lawrence 2001; Pampalk, Rauber, and Merkl 2002).

Rather than importing wholesale the cepstral coefficients developed for speech recognition (which, nonetheless, have performed well on music tasks in the past (Logan and Salomon 2001; Berenzweig, Ellis, and Lawrence 2003)), we will use a more general model of auditory preprocessing. As part of this, we have begun investigating subband modulation spectra (Kanedera, Hermansky, and Arai 1998) – the intensity of modulation in log-spaced bands between 2 and 30 Hz – as a helpful correlate of ‘smooth/rough’ in musical timbre.

3.1.2 Pitch/note transcription

As reviewed above, recovering the notes played in multi-voice music remains a very difficult task. We, however, believe that progress can be made by abandoning our ‘expert knowledge’ about musical notes (i.e., that the pitch is almost completely determined by the fundamental period of the waveform) used in all current approaches to transcription e.g. (Klapuri 2001; Goto 2001; Dannenberg and Hu 2002), and turning the problem over to trained classifiers. By freeing the computer of any preconceptions of the feature structure, we leave it able to take advantage of weaker statistical regularities, such as regular patterns of note co-occurrence, or subtle cues arising from the nonlinear interactions of polyphonic voices in the log-magnitude spectral domain.

The key factor in the balance between hand-designed and data-derived classifier is the availability of training data: Given enough training examples, a general machine learning algorithm ought to be able to infer any useful regularities. The classifier we wish to build takes a short time segment of polyphonic audio signal, and generates a set of posterior probabilities for each possible note/pitch (e.g. the 88 notes of a standard piano). Thus, the training data we need (for, e.g., a Multi-Layer Perceptron estimator, or family of Support Vector Machines) consists of musical recordings that are representative of the kind of music we wish to transcribe, along with the desired target outputs i.e. a temporally-aligned ‘piano-roll’ specifying exactly which notes are active at each time.

In general, such time-aligned target labels are not available for commercial music recordings. But by happy circumstance, there exists a large number of ‘imitations’ of popular music pieces encoded in MIDI format (including the pitch-timing information we need), created and made available on the internet by musician hobbyists. The best among these provide very faithful duplication of the original music – at least within the limited textural range of the General MIDI voices – meaning that most notes are present and in the correct order. The timing of these MIDI replicas does not perfectly line up with the original, since they are created independently rather than to “play along”, but if we could establish a single time-warping function between the MIDI replica and the original, we could map all the MIDI note-event times back through this function to obtain MIDI transcriptions corresponding precisely to the original commercial recording.

We have used dynamic time warp alignment, best known as a simple template-matching technique for speech recognition, to recover a high-accuracy time warping function between an audio resynthesis of MIDI replicas (whose timings exactly match the MIDI data), and the original recordings (Turetsky and Ellis 2003). (A similar approach has been proposed for the different task of finding transcripts to use in a query-by-humming system (Hu, Dannenberg, and Tzanetakis 2003).) When successful, these alignments give us exactly the data we are looking for – near-exact onset times and pitches for all the prominent note events in the original recordings. However, in our pilot study of 40 pieces, only 27 (68%) managed to make good-quality alignments. In the other cases, problems such as gross errors in transcription, omitted segments, and failures in our feature
normalization efforts to match MIDI and original acoustics, caused alignments that gave essentially random results, which would severely corrupt the training of our classifier. Thus, the limiting factor at this stage is a mechanism for reliably distinguishing between good and bad alignments. With several thousand MIDI replicas available, we can afford to be conservative in selecting only those whose alignments are very promising, and we are currently investigating accurate diagnostics for automatic rejection of poor alignments.

With these problems solved, we cantrain estimators of the likelihood of pitches present at each point in original music. This probabilistic note transcription provides one input to the subsequent music similarity and classification stage.

### 3.1.3 Chord transcription

Transcribing music into individual notes is difficult because an unknown number of notes will be present simultaneously; our best systems for classifying signals, such as speech recognizers, work by finding a single class label for each segment. If we could find a useful labeling of the music signal in terms of a succession of unique ‘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 one beat and several bars, but, critically, with only one chord present at a time. The many songs that share the same chord sequence have an immediately obvious ‘similarity’ if played alongside one another, and this factor should be available to our similarity and preference models.

Prior approaches to this problem have attempted to first identify notes, then form these into chords (Kashino, Nakadai, Kinoshita, and Tanaka 1998; Raphael 2002; Pardo and Birmingham 2001), and have had limited success. If the goal is to recover the chord identity, recovering the notes making up that chord is much more detail than required; it would be more parsimonious to recognize that chord on the basis of its common characteristics rather than 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, as with the note transcription, the precise time alignment of the chords to the sound waveforms is not known – only the sequence in which they will occur. This is directly analogous of the common problem in speech recognition, where the bulk of the training data may be transcribed into word sequences but without timing information within each utterance. 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/](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 songs. 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 songs achieved an (alignment) frame error rate of 16.7% when recognition was constrained to the correct chord sequence, but increased to 76.6% on unconstrained recognition (Sheh and Ellis 2003). These results confirm the viability of the approach and our feature set, and suggest that a substantially larger training set (which would not be too difficult to obtain) could give highly accurate transcripts.
### 3.1.4 Rhythm extraction

Various styles of music, including dances like the tango, obey strict constraints on their rhythmic structure, and such information should be considered in similarity and preference prediction. Drawing on the ideas of (Goto and Muraoka 1995), we are working on a system to recover a simplified percussion transcription using what amount to matched filters for two or three generic drum sounds, such as the ubiquitous bass drum, snare, and hi-hat of contemporary pop music. Starting from ‘cartoons’ of these instruments in the time-frequency magnitude domain, the matched-filter templates can be iteratively refined by averaging the best-quality matches to ‘tune in’ to the particular character of those instruments in different pieces. The templates then give soft estimates of the timing (and intensity) of each percussion event across the several instruments.

To generalize these patterns for matching to similar rhythms, we need to normalize out potential confounds such as tempo differences and errors in ‘downbeat’ alignment. Autocorrelation of the rhythm pattern estimates give the dominant period of repetition, and repeated cycles can then be averaged to give a single ‘master cycle’ characterizing the basic rhythm within the piece or segment. After normalizing the phrase durations, downbeats can be located by circular cross-correlation against a grand average drum pattern across the entire corpus (which again can be iteratively refined by averaging all normalized phrases in their new alignments, then repeating), or individual pairs of rhythm patterns can be compared by finding the maximum of their normalized circular cross-correlations.

### 3.2 Phrase segmentation

Almost every piece of music consists of a sequence of phrases or episodes, such as the alternating verse and chorus of the canonical pop song, with more or less different characteristics. Rather than making comparisons on the global average of the piece’s characteristics (as with previous work), we seek to first segment the piece into broadly consistent pieces, then match on the scale of each of these segments. Thus, the similarity between two songs with near-identical choruses could be detected, even if their contrasting choruses make them, on average, quite dissimilar.

Segmenting a continuous signal into sections that ‘belong together’ is a common problem, directly analogous to segmenting recorded discussions according to 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 the self-similarity matrix, a kind of Gram matrix giving distance comparisons between every pair of time frames in a piece (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).

To continue our theme of using classifiers derived from training data, we will also use a limited corpus of music with hand-labeled major phrase boundaries (produced by a high-school intern in
our lab last summer) as training material for a classifier looking at the kinds of features mentioned above to optimize the accuracy of segment boundary placement.

Given the estimated phrase segmentation, a piece of music is represented as a collection of its individual phrase-segments, with each segment represented by its features as described in section 3.1. These form the basis of the similarity calculations described in the next section.

3.3 Music comparison: Anchor space

Given the musically-relevant features describing each major phrase segment in a piece of music, we are now in a position to build our subjective similarity estimator. We do this through the concept of ‘anchor spaces’.

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 outputs (likelihoods or 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: \( \langle p(\omega_1|x), \ldots, p(\omega_M|x) \rangle \) where \( \omega_i \) represents the \( i \)th anchor 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, related to work on “supra-classifiers” and knowledge reuse (Bollacker and Ghosh 1998). Another related technique is the tandem acoustic modeling we have used 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 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 use the standard mel-frequency cepstral coefficients (MFCCs) from speech recognition.

The resulting attribute space was used to achieve 38% accuracy on a 400-class artist classification task (Berenzweig, Ellis, and Lawrence 2003), considerably more difficult than the 21-artist set which was the largest previously reported (Berenzweig, Ellis, and Lawrence 2002). (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 (Berenzweig, Logan, Ellis, and Whitman 2003) a centroid-based similarity measure in this space 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. In the best case, the choice of anchor labels should not influence the quality of the similarity space too greatly, provided the anchors provide a good spread across the range of material to be considered: it is irrelevant whether a particular piece scores a high match against any particular classifier, all that matters is that the signature of scores should change smoothly as the ‘nature’ of the music varies. (For this reason, we do not want any of the anchor classifiers to be too sharply discriminant; a slower falloff of likelihood over a broad segment of space is more desirable). For the full project, we will experiment with a much larger number of candidate dimensions and labels, looking for subjective quality attributes that are both successfully detectable using our base features at the same time as optimizing the coverage and independence of the attributes to minimize the dimensions required to cover the musical space. The work on automatic identification and grounding of musical-semantic terms through web searching (Whitman and Smaragdis 2002) may be useful.
Our pilot work in anchor space analysis has resulted in a prototype “music similarity browser”
interface, illustrated in figure 3 (Berenzweig, Ellis, and Lawrence 2003). The browser allows the
user to explore pieces of music similar to any given starting point by presenting its immediate
neighbors in anchor space. The user can also alter her location in anchor space by manipulating
the sliders to the right of the display, each bearing the semantic label estimated by that anchor
model. Thus, it is easy to search for “like Britney Spears, but with more soul”, at least to find what
the anchor model thinks this would sound like. The prototype can be accessed at www.playola.org.

The Playola prototype measures similarity as the overlap of ‘clouds’ of cepstral features, but
our more sophisticated musical features (pitch, harmony, and rhythm) require proportionally more
sophisticated similarity measures. As we discussed when describing the rhythm features,
simple normalizations (of duration, tempo, base key signature etc.) will be applied to simplify the
identification of similarity. The ‘fuzzy’ probabilistic nature of our transcriptions also aids in compa-

dison, since a ‘near miss’ is still likely to score a similarity greater than zero (through, e.g., tem-

plate cross-correlation). In general, however, it is impractical to enumerate all possible rhythms,
melodies, or chord sequences associated with a particular genre: some dimensionality reduction
is required. By normalizing a complex feature transcription onto a uniform grid, we can perform
principal component analysis (or variants like independent component analysis) to find ‘basis sets’
of eigenvectors that provide the best coverage of that dataset (according to whatever metric is
being optimized) within a reduced dimensionality. This kind of ‘eigenmelody’ (or eigenrhythm,
or eigen-chord-sequence) allows us to scale the dimensionality of the comparisons to match the
available ground-truth exemplar data.

3.4 Musical preference modeling

Having defined our perceptually-relevant similarity space, we return to the original motivation:
modeling musical preference. 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 similar, in this space, to items that the user is known to like. From this perspective, recommendation is simply information retrieval where the query is examples of ‘good’ music.

### 3.4.1 Listening behavior log analysis

In order to decide what the listener ‘likes’, we could just assume that a user’s entire personal music collection consists of preferred music. Finer grained distinctions, however, can be obtained by examining the records of actual listening behavior within that collection routinely created by today’s digital music devices (Apple’s iPod and its brethren). Not only can we identify which pieces of music are most preferred, based on number of plays, but we can also gain a deeper understanding of the time-course and variety of music preference experience, by seeing how different pieces have different profiles of play-frequency. We have been unable to find any existing longitudinal studies of this kind, presumably because of the huge problems in collecting this kind of data over months of listening, prior to digital music devices. We aim to make a passive, anonymous survey of listening habits by collecting the listening logs of volunteers drawn from the university community. Since the data is effortless to contribute, we are confident that we can collect dozens if not hundreds of long-term logs of this nature. This may also support breadth studies where we can compare the behavior of multiple listeners in respect of the same music, to see how much truth there is to the informal idea of different ‘kinds’ of tracks (quick hook, slow burn etc.) that work alike across all listeners.

Once we have identified a set of music that well describes a particular listener’s taste, the first step is to cluster that collection in anchor 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. For each cluster, the next step is to use principal component analysis (PCA) to find the dimensions of greatest and least variance. We then locally transform attribute space by the inverse of the PCA matrix to obtain a modified perceptual space where the cluster is close to spherical – i.e. dimensions with large variance in the original space, indicating relative indifference on the part of the listener, have been downweighted. In this space, simple Euclidean distance can be taken as an approximation of personalized preference-similarity. By projecting novel music into this space, and choosing the items closest to the cluster of preferred music, new songs to recommend are identified.

### 3.5 Evaluation

Evaluation of music similarity and recommendation systems is a particular challenge because, firstly, our essential goal of matching human judgments is subjective and obscure, and moreover it is highly variable between different users. Not only is there obvious source for ‘ground truth’ results in music or artist similarity, but there are serious questions over whether such a thing could even exist, or whether every listener defines her own ground truth. Bearing these challenges in mind, our evaluation plans are described below.

#### 3.5.1 Evaluating Features and Segmentation

The initial feature extraction parts of the project are amenable to independent evaluation since they are intended to detect human-defined aspects of the musical signal. Here, the MIDI-alignment
technique described in section 3.1.2 (which creates accurate note-level transcripts for real commercial recordings for which a high-quality MIDI replica can be found) can be used to obtain ground-truth data for quantitative measurement of system accuracy. We plan to hold out about 20 tracks from the hundreds of MIDI replicas we anticipate aligning, and these will be balanced to ensure a good coverage of different styles, instrumentations, etc. Additional hand marking will be performed for this test set to indicate the main structural elements (intro, verse, chorus, etc.), and this work has already begun as noted above. This many pieces constitutes over 100 segment boundaries, several hundred chords, and thousands of note events.

This data clearly provides the evaluation reference needed for the note recognition system, and this will be the first time that a polyphonic transcription algorithm has been quantitatively evaluated on this kind of real-world data, rather than artificial, synthetic test data. In addition, this data will support the evaluation of the chord recognition systems (through the assistance of MIDI-to-chord inference systems e.g. (Pardo and Birmingham 2001)), and the rhythm transcription module (by examination of the drum voices in the General MIDI vocabulary).

3.5.2 Evaluating Similarity Measures

We have recently focused on the problem of ground truth for music similarity systems (Ellis, Whitman, Berenzweig, and Lawrence 2002; Berenzweig, Logan, Ellis, and Whitman 2003). The essentially subjective nature of the task, and our research priority of quantitative evaluation, dictates that this is necessarily the starting point for any kind of work in music similarity, to avoid the fate of other researchers in this field who are often left only with subjective impressions, or a few suggestive examples, by which to gauge the quality of their work. In (Ellis, Whitman, Berenzweig, and Lawrence 2002), we constructed a web-based survey to collect direct subjective evaluations of artist similarities. To our amazement, this activity attracted the attention of various internet communities, and we ended up collecting over 22,000 trials from more than a thousand users. This data, combined with other online sources, allowed us to construct and evaluate a full similarity matrix, estimating an underlying ground-truth similarity between all pairs from a set of 400 contemporary pop music artists.

In (Berenzweig, Logan, Ellis, and Whitman 2003), we further developed mechanisms for using this data to evaluate signal-based similarity measures, including statistical significance measures. The net result of that paper was to find little or no significant difference between several different acoustic similarity measures, although they were clearly better than random guessing. More important, however, were the quantitative and systematic evaluation methodologies we developed, which we can continue to use as our similarity systems become more sophisticated.

We have made all this data available, in various easy-to-use formats, on our web site as a resource for other researchers in music similarity, in an effort to promote common evaluation standards (see http://www.ee.columbia.edu/ dpwe/research/musicsim/ ). A further obstacle to comparable evaluations is the copyright difficulties in sharing basic musical data; we have proposed a work-around in which we have offered to distribute our legally-owned collection of over 8000 music tracks, but only in the form of derived features (MFCCs, or whatever different researchers propose) so that the original music cannot be reconstructed with any fidelity (Logan, Ellis, and Berenzweig 2003).

3.5.3 Evaluating Preference Predictions/Recommendations

Music recommendation systems can be 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 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 personal collections are rarely constructed with the level of attention that this suggests.

Our analysis of digital music player logs, introduced in section 3.4.1 above, can overcome this problem. Given an extensive history of listening behavior, collected over months or even years of the project, we have a much finer indication of which tracks in a collection are truly preferred, and perhaps finer gradations of preference reflecting mood, sequential effects, etc. This longitudinal data will enable a very realistic test of the preference prediction by tracking the listener’s reaction to new music (such as new releases). We can train the preference model on data up until the time that the new music is first encountered, then gauge how well our system’s predicted preference ranking of the new material matches actual user behavior.

The most direct evaluation of the final system would be some kind of user study, and we will consider this possibility. The nature of music recommendation, however, is that it may take weeks or longer for a listener to decide how much they ‘really’ like a particular recommendation, so this data is slow and difficult to collect, and the chances of being able to demonstrate significant differences between alternative approaches are much worse than in the offline simulations described above.

4 Education, Outreach and Collaborations

Much of the pilot work described above has arisen from student projects conducted as part of the classes in Digital Signal Processing, and Speech and Audio Processing and Recognition, taught by the PI. Because of the intense appeal of music, excellent students will frequently invest disproportionate efforts in these projects. Funding to support a major research project in this area will provide many more opportunities for student projects that can tie in to, and leverage, the ongoing research.

Exciting and accessible results in music browsing will afford this project a high profile in a much broader community. The PI’s lab 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. We have also participated in a summer internship program for high-school students run by the New York Academy of Sciences, the source of the manual labels mentioned above.

We are in the planning stages of a collaborative project course run in conjunction with Columbia’s Computer Music Center (see attached letter of support from Computer Music Center director Prof. Brad Garton). This truly interdisciplinary effort would bring together interested engineering students with musicians and other artists to create installations and performances employing the latest computational techniques.

4.1 Workshop

As part of this project, we propose to organize a International Workshop on Music Signal Processing; this has been discussed with colleagues Prof. Mark Sandler of the University of London, and Dr. Masataka Goto of the National Institute of Advanced Industrial Science and Technology in Japan (see attached letters of support). Both these researchers work in highly related areas, and
agree that the time is ripe for a unifying workshop on topics of information abstraction from music signals. The workshop would be held probably in 2006 as a satellite to one of the more popular conferences such as the International Symposium on Music Information Retrieval (ISMIR) or the regular conventions of the Audio Engineering Society (AES). One idea for this workshop is to publish and promote a single set of annotated test material (as described in section 3.5.2 above) and make the use of this material (in some form or another) a condition of participation. While only semi-formal, this would be a nice way to help push the community toward common evaluation standards.

Given the thematic links between the proposed project and the new EU-funded SIMAC (Semantic Interaction with Music Audio Content) project at London, Prof. Sandler has also agreed to pursue the possibility of student exchanges between our labs, for instance over a summer. We have had success with similar exchanges with partners in the US and Europe before; they are an excellent way to cross-pollinate ideas, and are especially valued by the students involved.

4.2 Team

This work will be based at the Laboratory for Recognition and Organization of Speech and Audio (LabROSA), established by the PI 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 collaborate closely on this project, though separately funded through MIT (see attached letter from his adviser, Prof. Barry Vercoe). Dr. Beth Logan of HP Labs in Cambridge is a pioneer of content-based music analysis, and she has agreed to collaborate with us 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. Finally, after a successful student internship last summer, Google have expressed possible interest in content-based music browsing and retrieval, as expressed in their attached letter.

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, through our links with Columbia’s innovative Computer Music Center we can stay plugged-in to the vibrant musical culture of New York City.

5 Plan and Milestones

We anticipate specific work and achievements within the project to break down as follows:

Year 1: Initial development of novel music feature extractors, including creation of large note database from MIDI transcripts, expansion of chord recognition system data and models, and initial work on rhythm extraction; collection of manual ground-truth data for segmentation and listening logs/personal collection ground truth; initial analysis of listening log data to establish categories or patterns of music preference. Milestones: At least 100 tracks of high-quality MIDI transcription; chord recognition accuracy above 90% for general material; 3 months of listening history collected for at least 20 users.
Year 2: Development and evaluation of trained note transcription system; rhythm extraction system extended to produce comparable ‘normalized’ patterns; development of phrase segmentation system including self-similarity analysis and trained boundary detection; construction of anchor models based on music features, and investigation of different anchor choices; further analysis of listening log data. **Milestones:** 50% note error rate on transcription of real, commercial music; rhythm system achieves 80% agreement with human judgments of rhythm pattern equivalence; segmentation equal error rate below 20% in comparison with hand labels; anchor-based similarity model able to achieve top-N ranking agreement scores of 50% with survey ground truth (Berenzweig, Logan, Ellis, and Whitman 2003).

Year 3: Refinement of musical feature extractors and segmentation; improvement in anchor-based similarity models; development and testing of individualized preference models; integration into personalized music similarity browser prototype; possible user tests; organization of international workshop. **Milestones:** New music recommendations achieve 50% precision measured against listening log data; successful completion of workshop with at least 30 participants, and at least 6 groups reporting on a common data set.

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 combining audio structure analysis, geometric preference models, and inferred subjective ground-truth, we hope to explain significant components of musical preference, and at the same time develop new music search and browsing tools to unlock 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 a large number of analogous problems that occur everyday. Finally, the intrinsic interest and potential impact of our music-recommendation goal can serve as a positive ambassador for science and engineering throughout the whole population.

7 Results from prior support

PI Ellis is currently in the first 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). This project is concerned with using machine learning to recognize individual sources in sound mixtures. This year we have looked at different signal models for separating overlapped voices; one publication appeared at IEEE WASPAA-03 (Reyes-Gomez, Raj, and Ellis 2003), and a second paper is under review (Reyes-Gomez, Ellis, and Jojic).

Ellis is also a co-PI on NSF project IIS-0121396 “ITR/PE+SY: 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 is concerned with the application of speech recognition and other signal analysis techniques to information extraction from natural meetings. Publications from LabROSA include work on ‘episode’ segmentation (Renals and Ellis 2003), and finding emphasized utterances from pitch information (Kennedy and Ellis 2003).
References


tracking for audio signals. In *Working Notes of the IJCAI-95 Workshop on Computational
Auditory Scene Analysis*, pp. 68–75.


Herre, J., E. Allamance, and O. Hellmuth (2001). Robust matching of audio signals using spec-
tral flatness features. In *Proceedings of the 2001 IEEE Workshop on Applications of Signal
Processing to Audio and Acoustics*, Mohonk, New York, pp. 127–130.

Processing to Audio and Acoustics*, Mohonk NY.

trum for robust automatic speech recognition. In *Proceedings of IEEE International Con-

Kashino, K., K. Nakadai, T. Kinoshita, and H. Tanaka (1998). Application of the bayesian prob-
ability network to music scene analysis. In D. F. Rosenthal and H. Okuno (Eds.), *Compu-
com/kashino98application.html.

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

Klapuri, A. (2001). Multipitch estimation and sound separation by the spectral smoothness prin-

ACM Multimedia*.

Acoustics*, Mohonk, New York.


Logan, B., D. Ellis, and A. Berenzweig (2003, August). Toward evaluation techniques for music
SIGIR 2003*.

ICME 2001*, Tokyo, Japan.

Moorer, J. A. (1975). *On the Segmentation and Analysis of Continuous Musical Sound by Dig-


Dear Dan,

I'm writing to express my support for your proposal "Modeling Music Similarity: Signal-based Models of Subjective Preference". As you know, we have long been discussing ways of integrating our related work on music technology. The activities in your lab that this project would support would be ideal source material for the engineering/aesthetics projects class we have proposed, thereby providing benefit to a far wider group of students within the university.

I am deeply committed to developing activities that unite and integrate artistic fields with cutting-edge technology. Your work in applying pattern recognition to music signals is very exciting, and I would love to have the chance to work with you and your students more closely, to see what more unexpected uses we can put them to.

I very much hope that the NSF chooses to fund your proposal.

Brad Garton

Professor of Music
Director, Computer Music Center
Columbia University
January 7, 2004

Dear Prof. Dan Ellis:

I am writing regarding your project on “Modeling Music Similarity.” This is a very exciting project, and very relevant to my own work on extracting rhythmic and melodic descriptions from music signals. The research you propose will make an important and significant advance in the kinds of information that can be automatically derived from music recordings, opening up a whole new range of applications in smart services for music listeners and musicological analysis. I look forward to continuing our close contact, and I see many opportunities for informal collaboration in this area.

I am very glad that you are also proposing to organize a workshop on Music Signal Processing as part of the project, following on from my suggestions during our discussions in New York last fall. I see a very clear need for such a workshop: there must be a dozen or more groups around the world working on these topics, but in the absence of a clear common venue their work tends to be dissociated and poorly co-ordinated. At the same time, all these researchers have a great deal in common and I am sure that giving us a forum to discuss our areas of shared interest would be extremely productive. I would be delighted to be a co-organizer of this workshop, and I can liaise with the extensive music signal processing research community here in Japan to ensure there support.

Very best of luck with the proposal, and wishing you a happy new year,

Sincerely yours,

Masataka Goto

Dr. Masataka Goto
Information Technology Research Institute
National Institute of Advanced Industrial Science and Technology
Professor Dan P.W. Ellis  
Electrical Engineering  
Columbia University  
500 West 120th Street,  
New York, NY 10027  
USA

Dear Dan,

Concerning your proposal on Modeling Music Similarity, I am writing to express my strong support for the importance and significance of the work you are proposing. This is a very exciting time in content-based music analysis, with all kinds of new techniques coming into reach, and with many new applications waiting in the wings for the right kind of core enabling technologies. Your project should realize several of these. I particularly like your "semi-structured" approach of searching for explicit segments within each music piece, rather than relying on global average properties as in current content-based similarity systems. Although it is unrealistic to expect a complete automatic transcription in the near future, identifying and recovering a modest amount of the original musical structure strikes me as the appropriate balance in maximally expoting signal information to support human-relevant similarity models.

I also enthusiastically support your proposal for a workshop on music signal content processing, and I would be very happy to be a co-organizer with you. As you know, there is an extensive and active research community working on these topics in Europe at present, including my own group, IRCAM in Paris, UPF in Barcelona, and several others. I can use my close contacts with these groups to ensure that the workshop would be a success. One possibility would be to organize it as a satellite event at the 2005 International Conference on Music Information Retrieval which my group is hosting here in London. Another option is to hold it in association with one of the New York meetings of the Audio Engineering Society: it fits in nicely with the Semantic Audio Analysis activity I recently initiated with the AES, and which you are closely involved with.

This project would also open the door to much closer collaboration between our two groups. We are about to begin a closely-related project called SIMAC (Semantic Interaction with Music Audio Content) with funding from the European Union. Given the overlap in topic and in timescales of these two projects, I would be interested in encouraging exchanges of graduate students working on these projects for a period of a few months, perhaps over summers. This is an excellent way to cross-fertilize ideas, and has enormous broader educational benefits for the students involved.

Good luck to you in securing funding for this proposal, and I look forward to our future discussions and collaborations.

Very best wishes,

[Signature]

macintosh hd:users:mark:desktop:mark sandler headed.doc
Prof. Dan Ellis  
Department of Electrical Engineering  
Columbia University  
New York  
fax: (212) 932-9421

Dear Dan -

I am writing with regard to your proposal "Modeling Music Similarity: Automatic Prediction of Subjective Preference". I think this is an excellent proposal: I am enthusiastic about your exciting new ideas on using statistical methods to recover structure from musical recordings, and I am very pleased that some of the deep and long-standing questions of subjective musical preference will be addressed in this project.

I know my graduate student, Brian Whitman, has been working closely with you and your other students on this proposal and on several other recent projects in music analysis. This work is central to Brian’s dissertation research, and I will be very happy for him to continue this collaboration. Brian will continue to be funded through normal Media Lab channels.

I wish you the very best of luck with this proposal, and I dearly hope that you receive approval and support from the NSF. I look forward with hope to the prospect of collaborating with you, particularly through our shared involvement with Brian, on these topics in the future.

Warmest regards,

Barry Vercoe  
Professor of Media Arts & Sciences, MIT  
Founding Member, MIT Media Lab  
Assoc. Academic Head, MIT Media Lab
January 7 2004

Professor Dan Ellis
Columbia University
Mail Code 4712
500 West 120th Street, Room 1312
New York NY 10027

Dear Dan,

I wish to express my enthusiastic support for your proposal "Modeling Music Similarity: Signal-based Models of Subjective Preference". Developing new ways to find and organize music has become a crucial challenge since improvements in bandwidth and compression technology have brought closer a world in which all recorded music is ubiquitously available. An efficient mechanism for connecting consumers with musicians could change the very essence of music appreciation.

I am particularly impressed by your plan to develop a range of signal analyses to provide high-level descriptions of music in terms of spectral, harmonic, melodic, and rhythmic content to devise good similarity measures in those spaces. The search for the best features for this task is far from over and your work will contribute much to the field.

Hewlett-Packard recognizes this work as strategically significant, and, as you know, I have been working in related areas for several years. I am very excited by the prospect of you and your graduate students conducting a concerted effort on this topic, and would be delighted to continue to collaborate with you since we have a longstanding commitment to working with academia. I am currently spending up to 50% of my time working on music analysis and organization, and my hope would be that we could fully co-ordinate our work towards a single goal.

With best wishes

Dr. Beth Logan
Member of Research Staff
HP Labs
Cambridge MA
January 8, 2004

Dear Prof. Ellis,

I read your research proposal, "Modeling Music Similarity", with great interest. The issue of accurately predicting human subjective preference is a deep and challenging question, and a better ability to understand and model it in any domain could create exciting new applications.

Google is always considering new areas in which to expand our search and retrieval technologies, and content-based access to music is one intriguing possibility. We are happy to have established a relationship with your lab through the internship with Google of your student Adam Berenzweig, and we look forward to working with you more closely in the future. Certainly, if your project is able to find useful automatic methods for predicting when a user will 'like' even obscure music, and should Google decide to start an activity in this area, we would be very interested in exploiting the results of this research.

I would like to express my enthusiastic support for this project. I hope your proposal be viewed favorably by the National Science Foundation.

With best wishes,

Craig Nevill-Manning
Senior Research Scientist and Engineering Site Director, NYC
January 7, 2004

Dr. Dan Ellis
Department of Electrical Engineering
The Fu Foundation School of Engineering and Applied Science
Columbia University
New York, New York 10027

Dear Dr. Ellis:

I will be pleased to consult with you on the project, “Modeling music similarity: Automatic prediction of subjective preference,” which you are proposing to the National Science Foundation. The focus of the research is thoroughly and laudably interdisciplinary, and it is easy to see the potential of this work to inform several research communities about fundamental concerns and technical methods alike. I am especially eager to participate in devising the subjective measures of listeners to use as a standard for the modeling component of the project. Of course, I also agree to confer with you about the general cognitive dimensions of the research.

With my regards,

Robert E. Remez, Ph. D.
Ann Whitney Olin Professor
rer.8888