A Large-Scale Evaluation of Acoustic and Subjective Music Similarity Measures

Adam Berenzweig
LabROSA
Columbia University
New York NY U.S.A.
alb63@columbia.edu

Beth Logan
HP Labs
One Cambridge Center
Cambridge MA U.S.A.
beth.logan@hp.com

Daniel P.W. Ellis
LabROSA
Columbia University
New York NY U.S.A.
dpwe@ee.columbia.edu

Brian Whitman
Music Mind & Machine Group
MIT Media Lab
Cambridge MA U.S.A.
bwhitman@media.mit.edu

Abstract

Subjective similarity between musical pieces and artists is an elusive concept, but one that must be pursued in support of applications to provide automatic organization of large music collections. In this paper, we examine both *acoustic* and *subjective* approaches for calculating similarity between artists, comparing their performance on a common database of 400 popular artists. Specifically, we evaluate acoustic techniques based on Mel-frequency cepstral coefficients and an intermediate 'anchor space' of genre classification, and subjective techniques which use data from The All Music Guide, from a survey, from playlists and personal collections, and from web-text mining.

We find the following: (1) Acoustic-based measures can achieve agreement with ground truth data that is at least comparable to the internal agreement between different subjective sources. However, we observe significant differences between superficially similar distribution modeling and comparison techniques. (2) Subjective measures from diverse sources show reasonable agreement, with the measure derived from co-occurrence in personal music collections being the most reliable overall. (3) Our methodology for large-scale cross-site music similarity evaluations is practical and convenient, yielding directly comparable numbers for different approaches. In particular, we hope that our information-retrieval-based approach to scoring similarity measures, our paradigm of sharing common feature representations, and even our particular dataset of features for 400 artists, will be useful to other researchers.

Keywords: Music similarity, acoustic measures, evaluation, ground-truth.

1 Introduction

Techniques to automatically determine music similarity have attracted much attention in recent years (Ghias et al., 1995; Foote, 1997; Tzanetakis, 2002; Logan and Salomon, 2001; Aucouturier and Pachet, 2002; Ellis et al., 2002). Similarity is at the core of the classification and ranking algorithms needed to organize and recommend music. Such algorithms will be used in future systems to index vast audio repositories, and thus must rely on automatic analysis.

However, for the researcher or system builder looking to use similarity techniques, it is difficult to decide which is best suited for the task at hand. Few authors perform comparisons across multiple techniques, not least because there is no agreed-upon database for the community. Furthermore, even if a common database were available, it would still be a challenge to establish an associated ground truth, given the intrinsically subjective nature of music similarity.

The work reported in this paper started with a simple question: How do two existing audio-based music-similarity measures compare? This led us in several directions. Firstly, there are multiple aspects of each acoustic measure: the basic features used, the way that feature distributions are modeled, and the methods for calculating similarity between distribution models. In this paper, we investigate the influence of each of these factors.

To do that, however, we needed to be able to calculate a meaningful performance score for each possible variant. This basic question of evaluation brings us back to our earlier question of where to get ground truth (Ellis et al., 2002), and then how to use this ground truth to score a specific acoustic measure. Here, we consider five different sources of ground truth, all collected via the Web one way or another, and look at several different ways to score measures against them. We also compare them with one another in an effort to identify which measure is 'best' in the sense of approaching a consensus.

A final aspect of this work touches the question of sharing common evaluation standards, and computing comparable measures across different sites. Although common in fields such as speech recognition, we believe this is one of the first and largest cross-site evaluations in music information retrieval. Our work was conducted in two independent labs (LabROSA at Columbia, and HP Labs in Cambridge), yet by carefully specifying our evaluation metrics, and by sharing evaluation data in the form of derived features (which presents little threat to

copyright holders), we were able to make fine distinctions between algorithms running at each site. We see this as a powerful paradigm that we would like to encourage other researchers to use.

This paper is organized as follows. First we review prior work in music similarity. We then describe the various algorithms and data sources used in this paper. Next we describe our database and evaluation methodologies in detail. In Section 6 we discuss our experiments and results. Finally we present conclusions and suggestions for future directions.

2 Prior Work

Prior work in music similarity has focused on one of three areas: symbolic representations, acoustic properties, and subjective or 'cultural' information. We describe each of these below noting in particular their suitability for automatic systems.

Many researchers have studied the music similarity problem by analyzing symbolic representations such as MIDI music data, musical scores, and the like. A related technique is to use pitch-tracking to find a 'melody contour' for each piece of music. String matching techniques are then used to compare the transcriptions for each song e.g. (Ghias et al., 1995). However, techniques based on MIDI or scores are limited to music for which this data exists in electronic form, since only limited success has been achieved for pitch-tracking of arbitrary polyphonic music.

Acoustic approaches analyze the music content directly and thus can be applied to any music for which one has the audio. Blum et al. present an indexing system based on matching features such as pitch, loudness or Mel-frequency cepstral coefficients (MFCCs) (Blum et al., 1999). Foote has designed a music indexing system based on histograms of MFCC features derived from a discriminatively trained vector quantizer (Foote, 1997). Tzanetakis (2002) extracts a variety of features representing the spectrum, rhythm and chord changes and concatenates them into a single vector to determine similarity. Logan and Salomon (2001) and Aucouturier and Pachet (2002) model songs using local clustering of MFCC features, determining similarity by comparing the models. Berenzweig et al. (2003) uses a suite of pattern classifiers to map MFCCs into an "anchor space", in which probability models are fit and compared.

With the growth of the Web, techniques based on publicly-available data have emerged (Cohen and Fan, 2000; Ellis et al., 2002). These use text analysis and collaborative filtering techniques to combine data from many users to determine similarity. Since they are based on human opinion, these approaches capture many cultural and other intangible factors that are unlikely to be obtained from audio. The disadvantage of these techniques however is that they are only applicable to music for which a reasonable amount of reliable Web data is available. For new or undiscovered artists, an audio-based technique may be more suitable.

3 Acoustic Similarity

To determine similarity based solely on the audio content of the music, we use our previous techniques which fit a parametric probability model to points in an audio-derived input space (Logan and Salomon, 2001; Berenzweig et al., 2003). We then compute similarity using a measure that compares the models for two artists. The results of each measure are summarized in a *similarity matrix*, a square matrix where each entry gives the similarity between a particular pair of artists. The leading diagonal is, by definition, 1, which is the largest value.

The techniques studied are characterized by the features, models and distance measures used.

3.1 Feature Spaces

The feature space should compactly represent the audio, distilling musically important information and throwing away irrelevant noise. Although many features have been proposed, in this paper we concentrate on features derived from Mel-frequency cepstral coefficients (MFCCs). These features have been shown to give good performance for a variety of audio classification tasks and are favored by a number of groups working on audio similarity (Blum et al., 1999; Foote, 1997; Tzanetakis, 2002; Logan, 2000; Logan and Salomon, 2001; Aucouturier and Pachet, 2002; Berenzweig et al., 2003).

Mel-cepstra capture the short-time spectral shape, which carries important information about the music's instrumentation and its timbres, the quality of a singer's voice, and production effects. However, as a purely local feature calculated over a window of tens of milliseconds, they do not capture information about melody, rhythm or long-term song structure.

We also examine features in an 'anchor space' derived from MFCC features. The anchor space technique is inspired by a folk-wisdom approach to music similarity in which people describe artists by statements such as, "Jeff Buckley sounds like Van Morrison meets Led Zeppelin, but more folky". Here, musically-meaningful categories and well-known anchor artists serve as convenient reference points for describing salient features of the music. This approach is mirrored in the anchor space technique with classifiers trained to recognize musically-meaningful categories. Music is "described" in terms of these categories by running the audio through each classifier, with the outputs forming the activation or likelihood of the category.

For this paper, we used neural networks as anchor model pattern classifiers. Specifically, we trained a 12-class network to discriminate between 12 genres and two two-class networks to recognize these supplemental classes: Male/Female (gender of the vocalist), and Lo/Hi fidelity. Further details about the choice of anchors and the training technique are available in (Berenzweig et al., 2003). An important point to note is that the input to the classifiers is a large vector consisting of 5 frames of MFCC vectors plus deltas. This gives the network some time-dependent information from which it can learn about rhythm and tempo, at least on the scale of a few hundred milliseconds.

3.2 Modeling and Comparing Distributions

Because feature vectors are computed from short 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 model that distribution using statistical techniques. Extending this idea, we can conceive of a distribution in feature space that characterizes the entire repertoire of each artist.

We use Gaussian Mixture Models (GMMs) to model these

distributions, similar to previous work (Logan and Salomon, 2001). Two methods of training the models were used: (1) simple K-means clustering of the data points to form clusters that were then each fit with a Gaussian component, to make a Gaussian mixture model (GMM), and (2) standard Expectation-Maximization (EM) re-estimation of the GMM parameters initialized from the K-means clustering. Although unconventional, the use of K-means to train GMMs *without* a subsequent stage of EM re-estimation was discovered to be both efficient and useful for song-level similarity measurement in previous work (Logan and Salomon, 2001).

The parameters for these models are the mean, covariance and weight of each cluster. In some experiments, we used a single covariance to describe all the clusters. This is sometimes referred to as a "pooled" covariance in the field of speech recognition; an "independent" covariance model estimates separate covariance matrices for each cluster, allowing each to take on an individual 'shape' in feature space, but requiring many more parameters to be estimated from the data.

Having fit models to the data, we calculate similarity by comparing the models. The Kullback-Leibler divergence or relative entropy is the natural way to define distance between probability distributions. However, for GMMs, no closed form for the KL-divergence is known. We explore several alternatives and approximations: the "centroid distance" (Euclidean distance between the overall means), the Earth-Mover's distance (EMD) (Rubner et al., 1998) (which calculates the cost of 'moving' probability mass between mixture components to make them equivalent), and the Asymptotic Likelihood Approximation (ALA) to the KL-divergence between GMMs (Vasconcelos, 2001) (which segments feature space and assumes only one Gaussian component dominates in each region). Another possibility would be to compute the likelihood of one model given points sampled from the second (Aucouturier and Pachet, 2002), but as this is very computationally expensive for large datasets it was not attempted. Computationally, the centroid distance is the cheapest of our methods and the EMD the most expensive.

4 Subjective similarity measures

An alternative approach to music similarity is to use sources of human opinion, for instance by mining the Web. Although these methods cannot easily be used on new music because they require observations of humans interacting with the music, they can uncover subtle relationships that may be difficult to detect from the audio signal. Subjective measures are also valuable as ground truth against which to evaluate acoustic measures—even a sparse ground truth can help validate a more comprehensive acoustic measure. Like the acoustic measures, subjective similarity information can also be represented as a similarity matrix, where the values in each row give the relative similarity between every artist and one target.

4.1 Survey

The most straightforward way to gather human similarity judgments is to explicitly ask for it in a survey. We have previously constructed a website, musicseer.com, to conduct such a survey (Ellis et al., 2002). We defined a set of some 400 popular artists (described in section 5.3 below), then presented subjects with a list of 10 artists $(a_1, ..a_{10})$, and a single target artist a_t, and asked "Which of these artists is most similar to the target artist?" We interpret 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 *if* those artists are known to the subject, which we can infer by seeing if the subject has ever selected the artists in any context.

Ideally, the survey would provide enough data to derive a full similarity matrix, for example by counting how many times users selected artist a_i being most similar to artist a_j. However, even with the 22,000 responses collected, the coverage of our modest artist set is relatively sparse: only around 7.5% of all our artist pairs were directly compared, and only 1.7% of artist pairs were ever chosen as most similar. We constructed this sparse similarity matrix by populating each row with the number of times a given artist was chosen as most similar to a target as a proportion of the trials in which it could have been chosen. Although heuristic, this worked quite well for our data.

4.2 Expert Opinion

Rather than surveying the masses, we can ask a few experts. Several music-related online services contain music taxonomies and articles containing similarity data. The All Music Guide (www.allmusic.com) is such a service in which professional editors write brief descriptions of a large number of popular musical artists, often including a list of similar artists. We extracted the "similar artists" lists from the All Music Guide for the 400 artists in our set, discarding any artists from outside the set, resulting in an average of 5.4 similar artists per list (so 1.35% of artist pairs had direct links). 26 of our artists had no neighbors from within the set.

As in (Ellis et al., 2002) we convert these descriptions of the immediate neighborhood of each artist into a similarity matrix by computing the path length between each artist in the graph where nodes are artists and there is an edge between two artists if the All Music editors consider them similar. Our construction is symmetric, since links between artists were treated as nondirectional. We call this the Erdös measure, after the technique used among mathematicians to gauge their relationship to Paul Erdös . This extends the similarity measure to cover 87.4% of artist pairs.

4.3 Playlist Co-occurrence

Another source of human opinion about music similarity is human-authored playlists. We assume that such playlists contain similar music, which, though crude, proves useful. We gathered around 29,000 playlists from "The Art of the Mix" (www.artofthemix.org), a website that serves as a repository and community center for playlist hobbyists.

To convert this data into a similarity matrix, we start with the normalized playlist co-occurrence matrix, where entry (i, j) represents the joint probability that artist a_i and a_j occur in the same playlist. However, this probability is influenced by overall artist popularity which should not affect a similarity measure. Therefore, we use a normalized conditional probability matrix instead: Entry (i, j) of the normalized conditional probability matrix C is the conditional probability $p(a_i|a_j)$ divided by the prior probability $p(a_i)$. Since

$$c_{ij} = \frac{p(a_i|a_j)}{p(a_i)} = \frac{p(a_i, a_j)}{p(a_i)p(a_j)}, \quad (1)$$

this is an appropriate normalization of the joint probability. Note that the expected log of this measure is the mutual information $I(a_i; a_j)$ between artist a_i and a_j.

Using the playlists gathered from Art of the Mix, we constructed a similarity matrix with 51.4% coverage for our artist set (i.e. more than half of the matrix cells were nonzero).

4.4 User Collections

Similar to user-authored playlists, individual music collections are another source of music similarity often available on the Web. Mirroring the ideas that underly collaborative filtering, we assume that artists co-occurring in someone's collection have a better-than-average chance of being similar, which increases with the number of co-occurrences observed.

We retrieved user collection data from OpenNap, a popular music sharing service, although we were careful not download any audio files. After discarding artists not in our data set, we were left with about 176,000 user-to-artist relations from about 3,200 user collections. To turn this data into a similarity matrix, we use the same normalized conditional probability technique as for playlists as described above. This returned a similarity matrix nonzero values for 95.6% of the artist pairs.

4.5 Webtext

A rich source of information resides in text documents that describe or discuss music. Using techniques from the Information Retrieval (IR) community, we derive artist similarity measures from documents returned from Web searches (Whitman and Lawrence, 2002). The best-performing similarity matrix from that study, derived from bigram phrases, is used here. This matrix has essentially full coverage.

5 Evaluation Methods

In this section, we describe our evaluation methodology, which relies on some kind of ground truth against which to compare candidate measures; we expect the subjective data described above to be a good source of ground truth since they are derived from human choices. In this section we present several ways to use this data to evaluate our acoustic-based models, although the techniques can be used to evaluate any measure expressed as a similarity matrix. The first technique is a general method by which one can use one similarity matrix as a reference to evaluate any other, whereas the other techniques are specific to our survey data.

5.1 Evaluation against a reference similarity matrix

If we can establish one similarity metric as ground truth, how can we calculate the agreement achieved by other similarity matrices? We use an approach inspired by practice in text information retrieval (Breese et al., 1998): Each matrix row is sorted into decreasing similarity, and treated as the results of a query for the corresponding target artist. The top N 'hits' from the reference matrix define the ground truth, with exponentially-decaying weights so that the top hit has weight 1, the second hit has weight α_r, the next α_r^2 etc. (We consider only N hits to minimize issues arising from similarity information sparsity.) The candidate matrix 'query' is scored by summing the weights of the hits by another exponentially-decaying factor, so that a ground-truth hit placed at rank r is scaled by α_c^r. Thus this "top-N ranking agreement score" s_i for row i is

$$s_i = \sum_{r=1}^{N} \alpha_r^r \alpha_c^{k_r} \quad (2)$$

where k_r is the ranking according to the candidate measure of the r^{th}-ranked hit under the ground truth. α_c and α_r govern how sensitive the metric is to ordering under the candidate and reference measures respectively. With $N = 10$, $\alpha_r = 0.5^{1/3}$ and $\alpha_c = \alpha_r^2$ (the values we used, biased to emphasize when the top few ground-truth hits appear somewhere near the top of the candidate response), the best possible score of 0.999 is achieved when the top 10 ground truth hits are returned in the same order by the candidate matrix. Finally, the overall score for the experimental similarity measure is the average of the normalized row scores $S = \frac{1}{N}\sum_i^N s_i/s_{max}$, where s_{max} is the best possible score. Thus a larger rank agreement score is better, with 1.0 indicating perfect agreement.

One issue with this measure arises from the handling of ties. Because much of the subjective information is based on counts, ranking ties are not uncommon (an extreme case being the 26 'disconnected' artists in the "expert" measure, who must be treated as uniformly dissimilar to all artists). We handle this by calculating an average score over multiple random permutations of the equivalently-ranked entities; because of the interaction with the top-N selection, a closed-form solution has eluded us. The number of repetitions was based on empirical observations of the variation in successive estimates in order to obtain a stable estimate of the underlying mean.

5.2 Evaluating against survey data

The similarity data collected using our Web-based survey can be argued to be a good independent measure of ground truth artist similarity since users were explicitly asked to indicate similarity. However, the coverage of the similarity matrix derived from the survey data is only around 1.7%, which makes it suspect for use as a ground truth reference as described in section 5.1 above. Instead, we can compare the individual user judgments from the survey directly to the metric that we wish to evaluate. That is, we ask the similarity metric the same questions that we asked the users and compute an average agreement score.

We used two variants of this idea. The first, "average response rank", determines the average rank of the artists chosen from the list of 10 presented in the survey according to the experimental metric. For example, if the experimental metric agrees perfectly with the human subject, then the ranking of the chosen artist will be 1 in every case, while a random ordering of the artists would produce an average ranking of 5.5. In practice, the ideal score of 1.0 is not possible because survey subjects did not always agree about artist similarity; therefore, a ceiling exists corresponding to the single, consistent metric that optimally matches the survey data. For our data, this was estimated to give a score of 2.13.

The second approach is simply to count how many times the similarity measure agrees with the user about the first-place (most similar) artist from the list. This "first place agreement" proportion has the advantage that it can be viewed as the average of a set of independent binomial (binary-valued) trials, meaning that we can use a standard statistical significance test to confirm that certain variations in values for this measure arise from gen-

uine differences in performance, rather than random variations in the measure. Our estimate of the best possible first place agreement with the survey data was 53.5%.

5.3 Evaluation database

In order to conduct experiments we have compiled a large dataset from audio and Web sources. The dataset covers 400 artists chosen to have the maximal overlap of the user collection (OpenNap) and playlist (Art of the Mix) data. We had previously purchased audio corresponding to the most popular OpenNap artists and had also used these artists to construct the survey data. For each artist, our database contains audio, survey responses, expert opinions from All Music Guide, playlist information, OpenNap collection data, and webtext data.

The audio data consists of 707 albums and 8772 songs for an average of 22 songs per artist. Because our audio experiments were conducted at two sites, a level of discipline was required when setting up the data. We shared MFCC features rather than raw audio, both to save bandwidth and to avoid copyright problems. This had the added advantage of ensuring both sites started with the same features when conducting experiments. We believe that this technique of establishing common feature calculation tools, then sharing common feature sets, could be useful for future cross-group collaborations and should be seriously considered by those proposing audio music evaluations, and we would be interested in sharing our derived features. Duplicated tests on a small subset of the data were used to verify the equivalence of our processing and scoring schemes.

The specific track listings for this database, which we refer to as "uspop2002", are available at `http://www.ee.columbia.edu/~dpwe/research/musicsim/`.

6 Experiments and Results

A number of experiments were conducted to answer the following questions about acoustic- and subjective-based similarity measures:

1. Is anchor space better for measuring similarity than MFCC space?
2. Which method of modeling and comparing feature distributions is best?
3. Which subjective similarity measure provides the best ground truth, e.g. in terms of agreeing best, on average, with the other measures?

Although it risks circularity to define the best ground truth as the measure which agrees best with the others, we argue that since the various measures are constructed from diverse data sources and methods, any correlation between them should reflect a true underlying consensus among the people who generated the data. A measure consistent with all these sources must reflect the 'real' ground truth.

6.1 Acoustic similarity measures

We first compare the acoustic-based similarity measures, examining artist models trained on MFCC and anchor space features. Each model is trained using features calculated from the available audio for that artist. Our MFCC features are 20-dimensional and are computed using 32 ms frames overlapped by 16 ms. The anchor space features have 14 dimensions where each dimension represents the posterior probability of a prelearned acoustic class given the observed audio as described in Section 3.1.

In a preliminary experiment, we performed dimensionality reduction on the MFCC space by taking the first 14 dimensions of a PCA analysis and compared results with the original 20-dimensional MFCC space. There was no appreciable difference in results, confirming that any difference between the anchor-based and MFCC-based models is not due to the difference in dimensionality.

Table 1 shows results for similarity measures based on MFCC space, in which we compare the effect of varying the distribution models and the distribution similarity method. For the GMM distribution models, we vary the number of mixtures, use pooled or independent variance models, and train using either plain K-means, or K-means followed by EM re-estimation. Distributions are compared using centroid distance, ALA or EMD (as described in section 3.2). We also compare the effect of including or excluding the first cepstral coefficient, c_0, which measures the overall intensity of a signal. Table 1 shows the average response rank and first place agreement percentage for each approach.

From this table, we see that the different training techniques for GMMs give comparable performance and that more mixture components help up to a point. Pooling the data to train the covariance matrices is useful as has been shown in speech recognition since it allows for more robust covariance parameter estimates. Omitting the first cepstral coefficient gives better results, possibly because similarity is more related to spectral shape than overall signal energy, although this improvement is less pronounced when pooled covariances are used. The best system is one which uses pooled covariances and ignores c_0. Models trained with the simpler K-means procedure appear to suffer no loss, and thus are preferred.

A similar table was constructed for anchor-space-based methods, which revealed that full, independent covariance using all 14 dimensions was the best-performing method. Curiously, while the ALA distance measure performed poorly on MFCC-based models, it performed competitively with EMD on anchor space models. We are still investigating the cause; perhaps it is because the assumptions behind the asymptotic likelihood approximation do not hold in MFCC space.

The comparison of the best-performing MFCC and anchor space models is shown in Table 2. We see that both have similar performance under these metrics, despite the prior information encoded in the anchors.

6.2 Comparing ground truth measures

Now we turn to a comparison of the acoustic and subjective measures. We take the best-performing approaches in each feature space class (MFCC and anchor space, limiting both to 16 GMM components for parity) and evaluate them against each of the subjective measures. At the same time, we evaluate each of the subjective measures against each other. The results are presented in Table 3. Rows represent similarity measures being evaluated, and the columns give results treating each of our five subjective similarity metrics as ground truth. Top-N ranking agreement Scores are computed as described in section 5.1.

The mean down each column, excluding the self-reference diagonal, are also shown (denoted "mean*"). The column means can be taken as a measure of how well each measure approaches ground truth by agreeing with all the data. By this standard, the survey-derived similarity matrix is best, but its very sparse coverage makes it less useful. The user collection (opennap) data has the second-highest "mean*", including particularly high agreement with the survey metric, as can be seen when the top-N ranking agreements are plotted as an image in figure 1. Thus, we consider the user collections as the best source of a ground truth similarity matrix based on this evidence, with the survey (and hence the first place agreement metric) also providing reliable data. (Interestingly, the collection data does less well agreeing with the survey data when measured by the first place agreement percentage; we infer that it is doing better at matching further down the rankings).

We mentioned that a key advantage of the first place agreement measure was that it allowed the use of established statistical significance tests. Using a one-tailed test under a binomial assumption, first place agreements differing by more than about 1% are significant at the 5% level for this data (10,884 trials). Thus all the subjective measures are showing significantly different results, although differences among the variants in modeling schemes from tables 1 and 2 are at the edge of significance.

7 Conclusions and Future Work

Returning to the three questions posed in the previous section, based on the results just shown, we conclude:

1. MFCC and anchor space achieve comparable results on the survey data.

2. K-means training is comparable to EM training. Using pooled, diagonal covariance matrices is beneficial for MFCC space, but in general the best modeling scheme and comparison method depend on the feature space being modeled.

3. The measure derived from co-occurrence in personal music collections is the most useful ground truth, although some way of combining the information from different source warrants investigation since they are providing different information.

The work covered by this paper suggests many directions for future research. Although the acoustic measures achieved respectable performance, there is still much room for improvement. One glaring weakness of our current features is their failure to capture any temporal structure information, although it is interesting to see how far we can get based on this limited representation.

Based on our cross-site experience, we feel that this work points the way to practical music similarity system evaluations that can even be carried out on the same database, and that the serious obstacles to sharing or distributing large music collections can be avoided by transferring only derived features (which should also reduce bandwidth requirements). To this end, we have set up a web site giving full details of our ground truth and evaluation data, `http://www.ee.columbia.edu/~dpwe/research/musicsim/`. We will also share the MFCC features for the 8772 tracks we used in this work by burning DVDs to send to interested researchers. We are also interested in proposals for other features that it would be valuable to calculate for this data set.

Acknowledgments

We are grateful for support for this work received from NEC Laboratories America, Inc. We also thank the anonymous reviewers for their useful comments.

Much of the content of this paper also appears in our white paper presented at the Workshop on the Evaluation of Music Information Retrieval (MIR) Systems at SIGIR-03, Toronto, August 2003.

References

Aucouturier, J.-J. and Pachet, F. (2002). Music similarity measures: What's the use? In *International Symposium on Music Information Retrieval*.

Berenzweig, A., Ellis, D. P. W., and Lawrence, S. (2003). Anchor space for classification and similarity measurement of music. In *ICME 2003*.

Blum, T. L., Keislar, D. F., Wheaton, J. A., and Wold, E. H. (1999). *Method and article of manufacture for content-based analysis, storage, retrieval, and segmentation of audio information*. U.S. Patent 5, 918, 223.

Breese, J. S., Heckerman, D., and Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. In *Fourteenth Annual Conference on Uncertainty in Artificial Intelligence*, pages 43–52.

Cohen, W. W. and Fan, W. (2000). Web-collaborative filtering: recommending music by crawling the web. *WWW9 / Computer Networks*, 33(1-6):685–698.

Ellis, D. P., Whitman, B., Berenzweig, A., and Lawrence, S. (2002). The quest for ground truth in musical artist similarity.

Foote, J. T. (1997). Content-based retrieval of music and audio. In *SPIE*, pages 138–147.

Ghias, A., Logan, J., Chamberlin, D., and Smith, B. (1995). Query by humming. In *ACM Multimedia*.

Logan, B. (2000). Mel frequency cepstral coefficients for music modeling. In *International Symposium on Music Information Retrieval*.

Logan, B. and Salomon, A. (2001). A music similarity function based on signal analysis. In *ICME 2001*, Tokyo, Japan.

Rubner, Y., Tomasi, C., and Guibas, L. (1998). A metric for distributions with applications to image databases. In *Proc. ICCV*.

Tzanetakis, G. (2002). *Manipulation, Analysis, and Retrieval Systems for Audio Signals*. PhD thesis, Princeton University.

Vasconcelos, N. (2001). On the complexity of probabilistic image retrieval. In *ICCV'01*. Vancouver.

Whitman, B. and Lawrence, S. (2002). Inferring descriptions and similarity for music from community metadata. In *Proceedings of the 2002 International Computer Music Conference*. Sweden.

			Independent		Pooled		
	#mix	c0?	ALA	EMD	ALA	Cntrd	EMD
EM	8	y	4.76 / 16%	4.46 / 20%	4.72 / 17%	4.66 / 20%	4.30 / 21%
	8	n	-	4.37 / 22%	-	-	4.23 / 22%
	16	n	-	4.37 / 22%	-	-	4.21 / 21%
K-means	8	y	-	4.64 / 18%	-	-	4.30 / 22%
	8	n	4.70 / 16%	4.30 / 22%	4.76 / 17%	4.37 / 20%	4.28 / 21%
	16	y	-	4.75 / 18%	-	-	4.25 / 22%
	16	n	4.58 / 18%	4.25 / 22%	4.75 / 17%	4.37 / 20%	4.20 / 22%
	32	n	-	-	4.73 / 17%	4.37 / 20%	4.15 / 23%
	64	n	-	-	4.73 / 17%	4.37 / 20%	4.14 / 23%
Optimal			2.13 / 53.5%				
Random			5.50 / 11.4%				

Table 1: Average response rank / first place agreement percentage for various similarity schemes based on MFCC features. Lower values are better for average response rank, and larger percentages are better for first place agreement.

	MFCC	Anchor
#mix	EMD	ALA
8	4.28 / 21.3%	4.25 / 20.2%
16	4.20 / 22.2%	4.20 / 19.8%

Table 2: Best-in-class comparison of anchor vs. MFCC-based measures (average response rank / first place agreement percentage). MFCC system uses K-means training, pooled diagonal covariance matrices, and excludes c0. Anchor space system uses EM training, independent full covariance matrices, and includes c0.

	1st place	survey	expert	playlist	collection	webtext
Random	11.8%	0.015	0.020	0.015	0.017	0.012
Anchor	19.8%	0.092	0.095	0.117	0.097	0.041
MFCC	22.2%	0.112	0.099	0.142	0.116	0.046
Survey	53.5%	0.874	0.249	0.204	0.331	0.121
Expert	27.9%	0.267	0.710	0.193	0.182	0.077
Playlist	26.5%	0.222	0.186	0.985	0.226	0.075
Collection	23.2%	0.355	0.179	0.224	0.993	0.083
Webtext	18.5%	0.131	0.082	0.077	0.087	0.997
mean*		0.197	0.148	0.160	0.173	0.074

Table 3: First place agreement percentages (with survey data) and top-N ranking agreement scores (against each candidate ground truth) for acoustic and subjective similarity measures. "mean*" is the mean of each ground-truth column, excluding the shaded "cheating" diagonal and the "random" row.

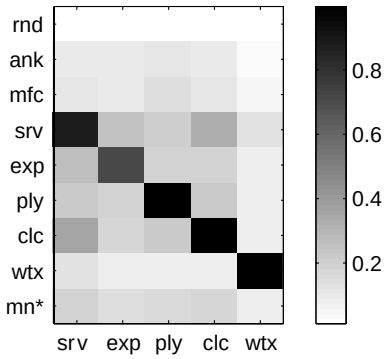

Figure 1: Top-N ranking agreement scores from table 3 plotted as a grayscale image.