Toward Evaluation Techniques for Music Similarity

Beth Logan, Daniel P.W. Ellis[1], Adam Berenzweig[1]
Cambridge Research Laboratory
HP Laboratories Cambridge
HPL-2003-159
July 29th , 2003*

E-mail: Beth.Logan@hp.com, dpwe@ee.columbia.edu, alb63@columbia.edu

music information retrieval, multimedia indexing, music analysis, information retrieval

We describe and discuss our recent work developing a database, methodology and ground truth for the evaluation of automatic techniques for music similarity. Our database consists of acoustic and textual 'Web-mined' data covering 400 popular artists. Of note is our technique of sharing acoustic features rather than raw audio to avoid copyright problems. Our evaluation methodology allows any data source to be regarded as ground truth and can highlight which measure forms the best collective ground truth. We additionally describe an evaluation methodology that is useful for data collected from people in the form of a survey about music similarity. We have successfully used our database and techniques to evaluate a number of music similarity algorithms.

* Internal Accession Date Only Approved for External Publication
[1] Columbia University, New York, NY
Published in and presented at SIGIR 2003: Workshop on the Evaluation of Music Information Retrieval Systems, 1 August 2003, Toronto, Canada

1 Introduction

The ubiquity of digital compression formats is transforming the way that people store, access and acquire music. Central to these changes is a need for algorithms to automatically organize vast audio repositories. Techniques to automatically determine music similarity will be a necessary component of such systems and as such have attracted much attention in recent years. [10, 9, 13, 11, 1, 8].

However, for the researcher or system builder looking to use or design similarity techniques, it is difficult to decide which is best suited for the task at hand simply by reading the literature. 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; It is not immediately clear how to obtain a reference ground truth for music similarity, since it is a naturally subjective phenomenon. It can vary not only across users, but across time, according to mood and according to context. Previous work has examined finding the ground truth for such a database [8].

In this paper, we describe our recently developed methodology and database for evaluating similarity measures. Our goal is to develop three key components necessary for a healthy community of comparable music similarity research: (1) A large scale, sharable database of features derived from real music; (2) ground truth results that best approach the ideal subjective outcomes, and (3) general, appropriate and accurate evaluation methodologies for this kind of work. Of these, the idea of a single ground truth is most problematic, since there is no particular reason to believe that similarity between two artists exists other than in the context of particular individual's taste. Although no two music listeners will completely agree, we still think it is useful to try and capture some kind of 'average' consensus.

We have previously validated our approach by comparing a variety of acoustic and subjective similarity measures on a large amount of common data at multiple sites [3]. Although our work has focused on artist similarity, our techniques extend to song similarity given a suitable database. We hope that our work will provide a helpful example and some useful techniques for other researchers to use. Ideally, we would like to see different sites contribute to a shared, common database of Web-mined features and copyright-friendly front-end features derived from their locally-owned music, as described below.

This paper is organized as follows. First we discuss some of the different kinds of music similarity measures in order to motivate the data and techniques required for evaluation. Next we describe our evaluation database, followed by the determination of ground truth and our evaluation methodologies. Finally, we discuss the results of our recent music similarity evaluation and our conclusions.

2 Music Similarity Measures

Music similarity measures rely on one of three types of information: symbolic representations, acoustic properties, and subjective or 'cultural' information. Let us consider each of these from the perspective of their suitability for automatic systems.

Many researchers have studied the music similarity problem by analyzing **symbolic representations** such as MIDI music data, musical scores, etc., or by using pitch-tracking to create a score-like 'melody contour' for a set of musical recordings. String matching techniques are then used to compare the transcriptions for each song. [4, 12, 10]. However, only a small subset of music has good-quality machine-readable score descriptions available, and automatic transcription becomes difficult and error-prone for anything other than monophonic music. Thus, pitch-based techniques are only applicable to single-voice music and approaches based on MIDI or scores can only be used for music which is already in symbolic form.

Acoustic approaches analyze the music content directly and thus can be applied to any music for which one has the audio. Most techniques use data derived from the short-term frequency spectrum and/or rhythm data. Typically, these features are modeled by one of a variety of machine learning techniques and comparisons in this domain are used to determine similarity [5, 9, 13, 11, 1, 2].

With the growth of the Web, techniques based on publicly-available data have emerged [7, 8, 14]. These use text analysis and collaborative filtering techniques to combine data from many individuals to determine similarity based on **subjective information**. 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, effective audio-based techniques would have a great advantage. Given our bias toward automatic techniques applicable to actual music recordings, we will focus on the latter two approaches in this paper. We now turn to the types of data required to determine similarity in the acoustic and 'web-mined' or subjective domains.

2.1 Data for Acoustic Similarity

Ideally, a database for evaluating acoustic similarity techniques would contain the raw audio of each song. This would enable an unlimited variety of features and models to be investigated and would additionally allow researchers to 'spot check' the results using their own judgment by listening to the pieces.

Unfortunately, copyright laws obstruct sharing data in this fashion. Until this issue is resolved (possibly a long wait), we propose instead the sharing of *acoustic features* calculated from the audio files. For example, in our recent evaluation we shared Mel-frequency cepstral coefficients (MFCCs) for each song. Starting from these common features, we were able to compare different algorithms on the same data, and we even saved some bandwidth transferring this data instead of the original waveforms. The best acoustic reconstruction possible from these reduced representations is only vaguely recognizable as the original music, so we are confident that sharing derived data of this kind will present no threat to copyright owners. Indeed, it is almost axiomatic that a good feature representation will eliminate much of the information present in the original signal, paring it down to leave only the essentials necessary for the task in question [1].

[1] Although it could be argued that subjective music similarity depends on practically all the information of interest to a listener, we confi dently predict that it will be many years before an automatic system attempts to make

MFCC features are currently popular as a basis for music similarity techniques, but their use is by no means as ubiquitous as it is in speech recognition. It is likely that over time researchers will add additional features to their repertoires. Until it is possible for sites to share raw audio then, we propose that authors share and distribute tools for the calculation of promising features. By downloading these tools and passing them over private collections, individual groups can generate features that can then be shared.

2.2 Data for Subjective Similarity

Subjective similarity can be determined using sources of human opinion mined from the Web. Here the required data is highly dependent on the technique used and the time at which the data was mined. We propose then that researchers using such techniques make their distilled datasets publicly available so that algorithms can be compared on the same data. We give examples of such datasets in the description of our database below.

3 Evaluation Database

Our database consists of audio and Web-mined data suitable for determining artist similarity. The dataset covers 400 artists chosen to have the maximal overlap of two of our main sources of Web-mined data: the artists best represented on the OpenNap peer-to-peer network in mid 2002, and the "Art of the Mix" playlist data from early 2003. We purchased audio and collected other data from the Web to cover these artists. We describe each of these sources in more detail below.

3.1 Audio Features

The audio data consists of 8827 songs with an average of 22 songs per artist. As described above, we pooled data between our different labs in the form of MFCC features rather than the original waveforms, 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.

3.2 Survey Data

Human similarity judgments came from our previously-constructed similarity survey website [8], which explicitly asked human informants for judgments: We defined a set of some 400 popular artists then presented subjects with a list of 10 artists $(a_1, ..a_{10})$, and a single target artist a_t, asking "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. For each subject. we infer which artists they know by seeing if the subject ever selects the artists in any context.

use of anything like this richness.

Ideally, the survey would provide enough data to derive a full similarity matrix, for example by counting how many times informants selected artist a_i being most similar to artist a_j. However, even with the 22,300 responses collected (from 1,000 subjects), the coverage of our modest artist set is relatively sparse.

3.3 Expert Opinion

Another source of data is expert opinion. 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 artist lists from the All Music Guide for the same 400 artists in our set, discarding any artists from outside the set, resulting in an average of 5.4 similar artists per list.

3.4 Playlist Co-occurrence

Yet another source of human opinion about music similarity is human-authored playlists. We assume that such playlists contain similar music — certainly an oversimplification, but one that turned out to be quite successful in our evaluations.
Again, the Web is a rich source for such playlists. In particular, we gathered over 29,000 playlists from "The Art of the Mix" , a website that serves as a repository and community center for playlist hobbyists (www.artofthemix.org). After filtering for our set of 400 artists, we were left with some 23,000 lists with an average of 4.4 entries.

3.5 OpenNap 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 did not download any audio files. After discarding artists not in our data set, we were left with about 175,000 user-to-artist relations from about 3,200 user collections.

3.6 Sparsity

A major difference between audio-based and subjective similarity measures lies in the area of data coverage: automatic measures based directly on the waveform can be applied to any pair of examples, even over quadratically-sized sets given sufficient computation time. Subjective ratings, however, inevitably provide *sparse* coverage, where only some subset of pairs of examples are directly compared. In the passive mining of subjective opinions provided by expert opinion and playlist and collection co-occurrence, there will be many artists who are never observed together, giving a similarity of zero. In the survey, we were able to choose which artists

Source	# obs	art/obs	> 0 obs	≥ 10 obs	med#art
Survey	17,104	5.54	7.49%	0.36%	23
Expert	400	5.41	1.35%	-	5
Playlist	23,111	4.38	51.4%	11.4%	213
Collection	3,245	54.3	94.1%	72.1%	388

Table 1: Sparsity of subjective measures: For each subjective data source we show the number of 'observations', the average number of valid artists in each observation, the proportion of the 79,800 artist pairs for which at least 1 co-occurrence or direct judgment was available, the proportion with 10 or more observations, and the median count of comparison artists (out of 400).

were presented for comparison, but even then we biased our collection in favor of choices that were believed to be more similar based on prior information. Specific sparsity proportions for the different subjective data sources are given in Table 1, which shows the proportion of all $400 \times 399/2$ artist pairs with nonzero comparisons/co-occurrences, the proportion with 10 or more observations (meaning estimates are relatively reliable), and the median number of artists for which some comparison information was available (out of 400). (For more details, see `http://www.ee.columbia.edu/~dpwe/research/musicsim/`.)

Two factors contribute to limit co-occurrence observations for certain artists. The first is that their subjective similarity may be very low. Although having zero observations means we cannot distinguish between several alternatives that are all highly dissimilar to a given target, this is not a particularly serious limitation, since making precise estimates of low similarity is not important in our applications. The second contributory factor, however, is unfamiliarity among the informant base: If very few playlists contain music by a certain (obscure) band, then we have almost no information about which other bands are similar. It is not that the obscure band is (necessarily) very different from most bands, but the 'threshold of dissimilarity' below which we can no longer distinguish comparison artists is much lower in these cases. The extreme case is the unknown band for which no subjective information is available – precisely the situation motivating our use of acoustic similarity measures.

4 Evaluation Methods

In this section, we describe our evaluation methodologies. The first technique is specific to the survey data which presents data in triplets and has sparse coverage. The second approach is a general way to compare two similarity matrices whose $(i, j)th$ element gives the similarity between artist i and artist j according to some measure. This technique is useful to gauge agreement between measures.

The choice of ground truth affects which technique is more appropriate. On the one hand, the survey explicitly asked subjects for similarity ratings and as such it might be regarded as a good source of ground truth. On the other hand, we expect many of the techniques based on the Web-mined data to be good sources of ground truth since they are derived from human choices.

4.1 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 subjects were explicitly asked to indicate similarity. We can compare the survey informant judgments directly to the similarity metric that we wish to evaluate. That is, we ask the similarity metric the same questions that we asked the subjects and compute an average agreement score.

We used two variants of this idea. The first, "average response rank", takes each list of artists presented to the informant and ranks it according to the similarity metric being tested. We then find the rank in this list of the choice picked by the informant (the 'right' answer), normalized to a range of 1 to 10 for lists that do not contain 10 items. The average of this ranking across all survey ground-truth judgment trials is the average response rank; 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 response rank of 5.5. In practice, the ideal score of 1.0 is not possible because informants do not always agree about artist similarity; therefore, a ceiling exists corresponding to the single, consistent metric that best matches the survey data. For our data, this was estimated to be 1.98.

A different way of using the survey data is to view each judgment as several 3-way subjudgments that the chosen artist a_c is more similar to the target a_t than each unchosen artist a_u in the list – that is

$$S(a_c, a_t) \geq S(a_u, a_t)$$

where $S(\cdot, \cdot)$ is the similarity metric. The "triplet agreement score" is computed by counting the fraction of such ordered "triplets" for which the experimental metric gives the same ordering.

4.2 Evaluation against similarity matrices

Although the survey data is a useful and independent evaluation set, it is in theory possible to regard any of our subjective data sources as ground-truth, and to seek to evaluate against them. Given a reference similarity matrix derived from any of these sources, we can use an approach inspired by the text information retrieval community [6] to score other similarity matrices. Here, each matrix row is sorted by decreasing similarity and treated as the result of a query for the corresponding target artist. The top N 'hits' from the reference matrix define the ground truth (where N is chosen to avoid the 'sparsity threshold' mentioned above) and are assigned exponentially-decaying weights so that the top hit has weight 1, the second hit has weight α_r, the next α_r^2 and so on, where $\alpha_r < 1$. The candidate similarity 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-1}. Thus this "top-N ranking agreement score" s_i for row i is

$$s_i = \sum_{r=1}^{N} \alpha_r^{r-1} \alpha_c^{k_r - 1}$$

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,

#mix	MFCC	Anchor
8	4.28 / 63%	4.25 / 64%
16	4.20 / 64%	4.19 / 64%
32	4.15 / 65%	-

Table 2: Survey evaluation metrics (average response rank / triplet agreement percentage) for K-means Models of MFCC features ('MFCC') and GMM models of Anchor Space features ('Anchor'). #mix gives the number of K-means clusters or mixture components.

biased to emphasize when the top few ground-truth hits appear somewhere near the top of the candidate response), the best possible score of 2.0 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 ranking agreement score is better, with 1.0 indicating perfect agreement.

5 Experimental Results

We have previously used our database and methodology to compare a variety of similarity measures [3]. These approaches succeeded in making possible comparisons between different parameter settings, models and techniques.

For example, Table 2 reproduces results from [3] comparing two acoustic-based similarity measures, using either a K-means cluster of MFCC features to model each artist's repertoire, compared via Earth-Mover's Distance [11], or a suite of pattern classifiers to map MFCCs into an "anchor space", in which probability models are fit and compared [2].

Table 2 shows the average response rank and triplets agreement score using the survey data as ground truth as described in Section 4.1. We see that both approaches have similar performance under these metrics, despite the prior information encoded in the anchors. It would have been very difficult to make such a close comparison without running experiments on a common database.

The scale of our experiment gives us confidence that we are seeing real effects. Access to a well-defined ground truth (in this case the survey data) enabled us to avoid performing user tests, which would have likely been impractical for this size database.

Using the techniques of Section 4.2 we were also able to make pairwise comparisons between all our subjective data measures, and to compare the two acoustic models against each subjective measure as a candidate ground truth. The rows in Table 3 represent similarity measures being evaluated, and the columns give results treating each of our five subjective similarity metrics as ground truth. Scores are computed as described in Section 4.2. For this scoring method, a random matrix scores 0.03 and the ceiling, representing perfect agreement with the reference, is 1.0.

Note the very high agreement between playlist and collection-based metrics: One is based on user-authored playlists, and the other on complete user collections. It is unsurprising that the

	survey	expert	playlist	collctn	mean*
survey	-	0.40	0.11	0.10	0.20
expert	0.27	-	0.09	0.07	0.14
playlst	0.19	0.23	-	0.58	0.33
collctn	0.14	0.16	0.59	-	0.30
Anchor	0.11	0.16	0.05	0.03	0.09
MFCC	0.13	0.16	0.06	0.04	0.10
mean*	0.17	0.21	0.16	0.15	

Table 3: Top-N ranking agreement scores for acoustic and subjective similarity measures with respect to each subjective measure as ground truth. "mean*" is the mean of the row or column, excluding the shaded "cheating" diagonal. A random ordering scores 0.03.

two agree. The moderate agreement between the survey and expert measures is also understandable, since in both cases humans are explicitly judging artist similarity. Finally, note that the performance of the acoustic measures is quite respectable, particularly when compared to the expert metric.

The mean down each row and column, excluding the self-reference diagonal, are also shown. We consider the row means to be an overall summary of the experimental metrics, and the column means to be a measure of how well each measure approaches as ground truth by agreeing with all the data. By this standard, the expert measure (derived from the All Music Guide) forms the best reference or ground truth.

6 Conclusions and Future Plans

We have described our recent work developing a database, methodology and ground truth for the evaluation of automatic techniques for music similarity. Our database covers 400 popular artists and contains acoustic and subjective data. Our evaluation methodologies can use as ground truth any data source that can be expressed as a (sparse) similarity matrix. However, we also propose a way of determining the 'best' collective ground truth as the experimental measure which agrees most often with other sources.

We believe our work represents not only one of the largest evaluations of its kind but also one of the first cross-group music similarity evaluations in which several research groups have evaluated their systems on the same data. Although this approach is common in other fields, it is rare in our community. Our hope is that we inspire other groups to use the same approach and also to create and contribute their own equivalent databases.

As such, we are open to adding new acoustic features and other data to our database. At present, we have fixed the artist set but if other sites can provide features and other data for additional artists these could be included. We would also welcome new feature calculation tools and scoring methodologies.

In order for this to take place, we are in the process of setting up a Website, www.musicseer.org, from which users can download our database, feature calculation tools and scoring scripts. Other

groups will be encouraged to submit their own data or features and scripts. We foresee no copyright problems given we are merely exchanging acoustic features that cannot be inverted into illegal copies of the original music. We hope that this will form the basis of a collective database which will greatly facilitate the development of music similarity algorithms.

7 Acknowledgments

Special thanks to Brian Whitman for the original OpenNap dataset, for help gathering the playlist data, and for generally helpful discussions.

References

[1] J-J Aucouturier and Francois Pachet. Music similarity measures: What's the use? In *Proc. Int. Symposium on Music Info. Retrieval (ISMIR)*, 2002.

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

[3] Adam Berenzweig, Beth Logan, Daniel P. W. Ellis, and Brian Whitman. A large-scale evaluation of acoustic and subjective music similarity measures. In *Submitted to Int. Symposium on Music Inform. Retrieval (ISMIR)*, 2003.

[4] S. Blackburn and D. De Roure. A tool for content based navigation of music. In *Proc. ACM Conf. on Multimedia*, 1998.

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

[6] John S. Breese, David Heckerman, and Carl Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Fourteenth Annual Conference on Uncertainty in Artificial Intelligence*, pages 43–52, 1998.

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

[8] Daniel P.W. Ellis, Brian Whitman, Adam Berenzweig, and Steve Lawrence. The quest for ground truth in musical artist similarity. In *Proc. Int. Symposium on Music Info. Retrieval (ISMIR)*, 2002.

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

[10] A. Ghias, J. Logan, D. Chamberlin, and B. Smith. Query by humming. In *ACM Multimedia*, 1995.

[11] Beth Logan and Ariel Salomon. A music similarity function based on signal analysis. In *ICME 2001*, Tokyo, Japan, 2001.

[12] R. McNab, L. Smith, I. Witten, C. Henderson, and S. Cunningham. Towards the digital music library: Tune retrieval from acoustic input. In *Digital Libraries 1996*, pages 11–18, 1996.

[13] G. Tzanetakis. *Manipulation, Analysis, and Retrieval Systems for Audio Signals*. PhD thesis, Princeton University, 2002.

[14] Brian Whitman and Steve Lawrence. Inferring descriptions and similarity for music from community metadata. In *Proc. Int. Comp. Music Conf.* Sweden, 2002.