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

Adam Berenzweig[1], Beth Logan[2], Daniel P.W. Ellis[1], Brian Whitman[3]
[1] LabROSA, Columbia University, New York USA
[2] HP Labs, Cambridge MA USA
[3] Music, Mind & Machine Group, MIT Media Lab, Cambridge MA USA

**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, and from playlists and personal collections.

We find the following: (1) Acoustic-based measures, when carefully constructed, can achieve high agreement with subjective measures. We report significant differences between superficially similar distribution modeling and comparison techniques. (2) Subjective measures from diverse sources also 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 multi-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 [8, 7, 12, 10, 1, 6]. 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. Our previous work was a first attempt at establishing ground truth for such a database [6].

In this paper, we describe a multi-site evaluation comparing a number of similarity techniques on a large common database. We focus on the problem of artist similarity, rather than considering individual songs or other units. Our aim is threefold. First, we wish to develop and verify good techniques for computing artist similarity. A basic question is whether similarity measures that are computed from the audio signal are comparable with measures derived from sources of subjective human opinion such as surveys, experts, playlists, user collections, and written documents. Acoustic-based measures are attractive because they can be computed for new or unpublicized music. However, music understanding is an extremely complex problem, so subjective data sources are likely to uncover relationships not easily discerned by machine.

Second, we continue our work toward developing an evaluation methodology for music similarity research. We propose a way to use subjective judgments about music similarity mined from the Internet as a collective soft ground truth. Similarity matrices derived from different sources will give different results, but the experimental measure which agrees most often with these varied sources can, in some sense, be claimed as the best measure.

Finally, 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 less so in our community. One outcome of this work is a common database which can be shared between groups as well as an evaluation methodology for using it. Our hope is that this paper inspires other groups to use the same approach and also to create and contribute their own equivalent databases.

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. [8]). 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) [3]. Foote has designed a music indexing system based on histograms of MFCC features derived from a discriminatively trained vector quantizer [7]. Tzanetakis extracts a variety of features representing the spectrum, rhythm and chord changes and concatenates them into a single vector to determine similarity [12]. Logan [10] and Aucouturier [1] model songs using local clustering of MFCC features, determining similarity by comparing the models. Berenzweig [2] 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 [5, 6]. 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 [10]. 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 [3, 7, 12, 9, 10, 1, 2].

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 [?]. 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 [10]. Two methods of training the models were used: (1) the standard Expectation Maximization (EM) algorithm for the GMM and (2) K-means clustering. Although unconventional, the use of K-means to train GMMs was discovered to be useful to measure song-level similarity in previous work [10].

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)[11], and the Asymptotic Likelihood Approximation (ALA) to the KL-divergence between GMMs [13]. Another possibility would be to compute the likelihood of one model given points sampled from the second [1], 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

Another approach to music similarity is to use sources of human opinion mined from 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 similarity measures are also represented in 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 [6]. 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 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 10,000 responses collected, the coverage of our modest artist set is relatively sparse.

## 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 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.

As in [6], 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.

## 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 is certainly an oversimplification, but as we will see it proves useful.

The Web is a rich source for such playlists. In particular, we gathered over 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. We would like to compute the 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 joint probability is influenced by overall artist popularity which should not effect a similarity measure. Therefore, we compute 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

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

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 29,000 playlists gathered from Art of the Mix, we constructed a similarity matrix with 51% 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 are more similar than randomly-chosen artists.

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 400,000 user-to-song relations from about 3,700 user collections. To turn this data into a similarity matrix, we use the same normalized conditional probability technique as for playlists as described above.

## 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 [14]. The best-performing similarity matrix from that study, derived from bigram phrases, is used here.

# 5 Evaluation Methods

In this section, we describe our evaluation methodology. A crucial decision here is the choice of ground truth. We expect the subjective measures 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 applied to any similarity matrix. The first technique is a general method by which one can use one similarity matrix as a reference to evaluate any other. We also present several other techniques that 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 [4]: 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 $\alpha_r$, the next $\alpha_r^2$ etc. 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 $\alpha_c^{r-1}$. Thus the 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. $\alpha_c$ and $\alpha_r$ 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 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 rank agreement score is better, with 1.0 indicating perfect agreement.

## 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 5%. Therefore, it is unclear whether this sparse matrix should be used as a ground truth reference as described in Section 5.1 above. Instead, we can compare the survey user 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 users and compute an average agreement score.

We used two variants of this idea. The first, "average rank agreement", 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 users 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.

The second approach is to view each user judgment as several 3-way sub-judgments 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 second evaluation score is computed by counting the fraction of such ordered "triples" for which the experimental metric gives the same ordering.

### 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 OpenNap and 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 then, our database contains audio, survey responses, expert opinions from All Muisc Guide, playlist information, OpenNap collection data and webtext data.

The audio data consists of 8827 songs with an average of 22 songs per artist. We conducted audio experiments at several sites, enforcing a level of discipline 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.

## 6 Experiments and Results

A number of experiments were conducted to answer the following questions about acoustic- and subjective-based similarity measures: (1) Which method of modeling and comparing feature distributions is best? (2) Is anchor space better for measuring similarity than MFCC space? (3) Which subjective similarity measure provides the best ground truth, that is, which agrees best on average with the other measures?

Although it seems circular 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 such diverse data sets 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 similiarty 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 32ms frames overlapped by 16ms. The anchor space features are of dimension 14 where each dimension represents the posterior probability of a pre-learned acoustic class given the observed audio as described in Section 3.1.

In a preliminary experiment, we performed dimensionality reduction on the MFCC space to reduce it to 14 dimensions 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. We show the average response rank and triples agreement score for MFCC features modeled by GMMs trained using standard EM and K-means modeling, and compared using various between-model distance measures. Results are shown for various numbers of mixture components, using pooled and non-pooled covariance matrices, and using or ignoring the first cepstral feature $c0$.

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 robustly trained covariance parameters. 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 c0. Also, since the K-means models are much cheaper to train than those trained using full EM without incurring an accuracy loss, they are preferable.

|  |  |  | Non-Pooled |  | Pooled |  |  |
|---|---|---|---|---|---|---|---|
|  | #mix | c0? | ALA | EMD | ALA | Cntrd | EMD |
| Ideal | - | - | 1.98/0.00 |  |  |  |  |
| EM | 8 | y | 4.76/0.42 | 4.46/0.39 | 4.72/0.41 | 4.66/0.41 | 4.30/0.37 |
|  | 8 | - | - | 4.37/0.37 | - | - | 4.23/0.36 |
|  | 16 | - | - | 4.37/0.37 | - | - | 4.21/0.36 |
| K-mn | 8 | y | - | 4.64/0.37 | - | - | 4.30/0.37 |
|  | 8 | - | 4.70/0.41 | 4.30/0.37 | 4.76/0.42 | 4.37/0.38 | 4.28/0.37 |
|  | 16 | y | - | 4.75/0.42 | - | - | 4.25/0.36 |
|  | 16 | - | 4.58/0.40 | 4.25/0.36 | 4.75/0.41 | 4.37/0.38 | 4.20/0.36 |
|  | 32 | - | - | - | 4.73/0.41 | 4.37/0.38 | 4.15/0.35 |
|  | 64 | - | - | - | 4.73/0.41 | 4.37/0.38 | 4.14/0.35 |

Table 1: Selected survey evaluation metrics for various similarity measures based on MFCC features. Metrics are: average response rank/triples error rate. Lower numbers indicate better similarity measures.

|  | MFCC | Anchor |
|---|---|---|
| #mix | EMD | ALA |
| 8 | 4.28/0.37 | 4.25/0.36 |
| 16 | 4.20/0.36 | 4.19/0.36 |
| 32 | 4.15/0.35 | - |

Table 2: Best-in-class comparison of anchor vs. MFCC-based measures. Survey evaluation metrics for MFCC (kmeans, pooled diag cov, no c0) and anchor (EM, unpooled full cov, c0).

A similar table was constructed for anchor-space-based methods, which revealed that full, unpooled 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 Cross-reference evaluation

Now we turn to a comparison of the acoustic and subjective measures. We take the best-performing measures in each feature space class (MFCC and anchor space) 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. Scores are computed as described in section 5.1. For this scoring method, a random matrix scores 0.07 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 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.

## 7 Conclusions and Future Work

Returning to the three questions posed in the previous section, based on the results just shown, we conclude: (1) K-means training is comparable to EM training. Using pooled, diagonal covariance matrices is beneficial for MFCC space. The best model comparison method depends on which feature space is being modeled. (2) MFCC and anchor space achieve comparable results on the survey data. (3) The measure derived from All Music Guide expert opinion is the best ground truth.

The work covered by this paper suggests many directions for future research. Although the acoustic measures achieved respectable performance there is still room for improvement, for instance by using a richer feature set with a longer temporal perspective. Also, we are excited by the promise of databases and evaluation methodologies that can be shared across groups, and we will work to develop this idea.

|  | survey | expert | playlist | collctn | webtext | mean* |
|---|---|---|---|---|---|---|
| Anchor | 0.11 | 0.16 | 0.05 | 0.03 | 0.04 | 0.08 |
| MFCC | 0.13 | 0.16 | 0.06 | 0.04 | 0.05 | 0.09 |
| survey | - | 0.40 | 0.11 | 0.10 | 0.12 | 0.18 |
| expert | 0.27 | - | 0.09 | 0.07 | 0.07 | 0.13 |
| playlst | 0.19 | 0.23 | - | 0.58 | 0.09 | 0.27 |
| collctn | 0.14 | 0.16 | 0.59 | - | 0.08 | 0.24 |
| webtext | 0.15 | 0.14 | 0.07 | 0.07 | - | 0.11 |
| mean* | 0.17 | 0.21 | 0.16 | 0.15 | 0.08 | 0.15 |

Table 3: 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 0.07.

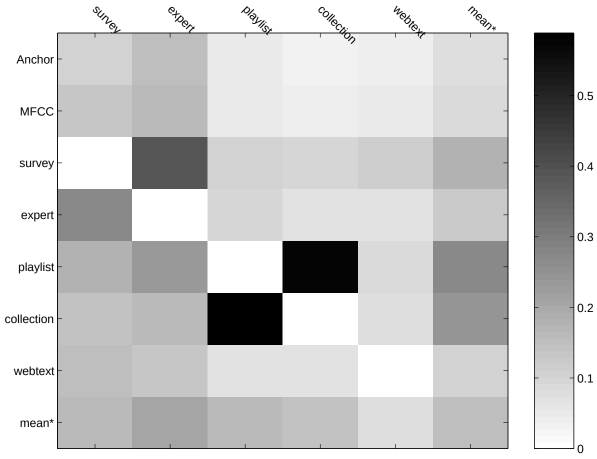

Figure 1: Table 3 plotted as a matrix.

# References

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

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

[3] 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.

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

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

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

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

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

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

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

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

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

[13] N. Vasconcelos. On the complexity of probabilistic image retrieval. In *ICCV'01.* Vancouver, 2001.

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