A STUDY ON ATTRIBUTE-BASED TAXONOMY FOR MUSIC INFORMATION RETRIEVAL

Jeremy Reed
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332
jeremy.reed@gatech.edu

Chin-Hui Lee
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332
chl@ece.gatech.edu

ABSTRACT

We propose an attribute-based taxonomy approach to providing alternative labels to music. Labels, such as genre, are often used as ground-truth for describing song similarity in music information retrieval (MIR) systems. A consistent labelling scheme is usually a key in determining quality of classifier learning in training and performance in testing of an MIR system. We examine links between conventional genre-based taxonomies and acoustical attributes available in text-based descriptions of songs. We show that the vector representation of each song based on these acoustic attributes enables a framework for unsupervised clustering of songs to produce alternative labels and quantitative measures of similarity between songs. Our experimental results demonstrate that this new set of labels are meaningful and classifiers based on these labels achieve similar or better results than those designed with existing genre-based labels.

1. INTRODUCTION

In recent years, the performance in genre recognition has reached an asymptotic maximum level of around 75-85% [1]. Many researchers have blamed ground-truth labelling procedures and debated the significance of the genre recognition task. Largely, two factors are often cited as limitations in designing reasonable training and testing databases for the task of genre recognition, and more generally, music similarity. The first is the inconsistent labelling of music, both in terms of how a particular artist or piece is classified and in the various class labels used in the labelling procedure [2]. In part, this can be explained by the tendency for people to use more finely grained categories for particular classes they enjoy [3]. For example, a listener of *contemporary* music may use ten to twenty different subclasses of *rock*, but may only have one general definition for *classical*. Meanwhile, a *classical* aficionado may use *romantic*, *baroque*, etc., but describe all forms of *rock*, *hip-hop*, etc. as *contemporary*.

The second factor affecting database design is that current genre labels are extracted from record companies and web-base retailers, who assign labels at the album or artist level rather than to individual songs [4]. Several questions arise because of this issue. If only one label is given per artist/album, does that really mean the entire collection is indeed similar? In the case that multiple labels are given to an artist, how does one decide which song has a particular label(s)? For instance, allmusic.com[1] has six styles for Eric Clapton, including *hard rock* and *adult contemporary*. Obviously, finding example songs that would fall under both categories is rare. This also demonstrates that by allowing fuzzy classification the database design issue is not solved; i.e., each label given in a multiple-class object still needs to be correctly applied.

Another problem in building databases for MIR purposes is that a person's concept of musical genre and musical similarity does not include just auditory cues [5]. For example, allmusic.com lists *heavy metal* and *thrash* as styles for Metallica's "Load" album, even though many die-hard *metal* fans criticized the album for its *alternative rock* leanings. On possible solution to deal with this issue is to incorporate non-acoustically-based algorithms, such as collaborative filtering [6]; however, there is still a desire to separate the testing and tuning of these two sub-systems. Further, collaborative filtering approaches can only be used when textual data exists and is therefore unavailable for new music. Another solution is to define genre labels based solely on music theory [7]. However, there is evidence that both musicians and non-musicians do not use such deep levels of musical understanding [8].

Given these issues, the previously used genre labels are inconsistent, lack precision, and are biased by non-acoustical cues. These issues have given some researchers the conclusion that genre is not a solvable problem and that instead, focus should be concerned with general notions of similarity. However, as pointed out by McKay and Fujinaga, the issues concerning genre also influence similarity performance metrics [4]. For example, if non-acoustic factors influence the perception of similarity, user studies must be able to isolate when a decision is based on non-acoustical information if the end result is to test auditory-based classifiers. In addition, there is plenty of literature demonstrating that

[1] http://allmusic.com

people use genre when searching for music [9] and that descriptions of genres are well-formed [3].

This paper proposes a novel construction for musical databases using clustering techniques on musical attributes, which are connected to specific acoustic qualities from either a surface-level (e.g., timbre) or a high-level (e.g., song form). By building taxonomy from a consistently-applied set of acoustic labels, non-acoustical information does not bias the performance metrics. Remaining songs can then be described in terms of their distances from each respective class, allowing for fuzzy classification that is based on a "relevance score" in terms of similarity of attributes. Section 2 of this paper describes the overall construction of the new labelling scheme and illustrates how acoustical properties guide the building of taxonomy. Section 3 describes how discriminative training reinforces cluster intra-similarity and increases the inter-cluster distances. Section 4 demonstrates the increased performance between the proposed unsupervised clustering and the traditional genre labels given at the artist level. Section 5 demonstrates that song-level genre labels still possess large variations in acoustic attributes. Conclusions and future considerations are given in Section 6.

2. CLUSTERING PROCEDURE

2.1. Taxonomy Construction

Due to inconsistencies found in previous genre labels, this section describes a new labelling procedure for producing ground-truth similarity labels. This new procedure is designed such that the following criteria are met:

- Labels are consistent: taxonomy labels are built from attributes that are assigned in a consistent manner by human subjects.
- Appropriate precision: taxonomies are meaningful in terms of precision and not too general (e.g., 2 genres: *classical* and *popular*)
- Acoustically meaningful: taxonomies for acoustically-based algorithms can only hope in learning information contained in the audio signal, whether at the surface level (e.g., timbre, rhythm) or at a high-level (e.g., song form, tonal progression).

In order to meet these requirements, previous genre labels are ignored and a new set of labels is developed in an unsupervised fashion using acoustically meaningful descriptors from the Music Genome Project[1] (MGP). The rationale behind using MGP information is that attributes given to each song are musically motivated and text-normalization, such as stemming and stop-word filtering [10], is easy given the small attribute set. The example in Figure 1 demonstrates that the attribute list can be seen as musical "subjects." The list presents several surface-level features such as rhythm and instrumentation (timbre). In addition, there are high-level acoustic features dealing with tonality and song structure.

basic rock song structures
a subtle use of vocal harmony
repetitive melodic phrasing
abstract lyrics
a twelve-eight time signature
extensive vamping
mixed acoustic and electric instrumentation
a vocal-centric aesthetic
major key tonality
electric guitar riffs
acoustic rhythm guitars
triple note feel

Figure 1. Example: Music Genome Project attributes.

There is information that is not relevant to many MIR acoustic-based similarity applications. For example, lyric identification is often seen as a separate task and not relevant for the majority of acoustic-similarity algorithms, which try link songs based on the musical qualities and not verbal subject matter. Fortunately, given the small attributes set provided by MGP, which is around 500 attributes, text normalization is an easy procedure. A musically meaningful word list was manually created from the original set of attributes, resulting in 375 words. Since this is an initial study and the authors felt that the creation of a "final" word list should be agreed upon by the MIR community, the tendency was to leave words that could be meaningful. For example, the word "off" was left in the list because "off beat" carries a very definite musical quality, even though "off" is a very commonly filtered word in text processing tasks. The authors also acknowledge that sources other than MGP might provide useful information. However, the source selection should insure the criteria stated here are meet.

2.2. Latent Semantic Analysis

Latent semantic analysis (LSA) [11] represents a document collection with a term-document matrix where rows correspond to individual terms and columns represent documents. By including bigrams, which is the appearance of two consecutive words, each column vector has size $M = J + J*J$, with J equal to the number of unigrams (i.e., number of words in the lexicon). Specifically, each element of the term-document matrix, W, is

$$w_{i,j} = (1 - \varepsilon_i)\frac{c_{i,j}}{n_j} \quad (1)$$

where $c_{i,j}$ is the number of times word i appears in document j and n_j is the word count of document j. The

[1] www.pandora.com

size of the matrix W is M x N, where N represents the number of documents. The term ε_i is the normalized entropy for term i and is given by

$$\varepsilon_i = -\frac{1}{\log N}\sum_{j=1}^{N}\frac{c_{i,j}}{t_i}\log t_i \qquad (2)$$

where t_i is the total number of times term i appears in the training database. The entropy gives a measure of indexing power, where a value close to one indicates the word appears in very few documents.

The resulting matrix is very sparse, and therefore, feature reduction is performed through singular value decomposition (SVD), which is very close to eigenvalue decomposition [12]. In SVD, W, is decomposed into

$$W \approx \hat{W} = USV^T \qquad (3)$$

where U is MxQ, S is QxQ, V is NxQ and Q is the rank of the original matrix, W. The left-singular matrix, U, and the right-singular matrix, V, represent the term and document space, respectively. The matrix S is a diagonal matrix of singular values, which represent the variation along the Q axes. By keeping $Q_0 \le Q$ singular values, the word-document space can be converted into a lower-dimensional "concept" space [11].

2.3. Song Clustering

The cosine similarity is a natural measure of distance and is given by

$$K(d_i, d_j) = \cos(v_i S, v_j S) = \frac{v_i S^2 v_j^T}{\|v_i S\|\|v_j S\|} \qquad (5)$$

where $\|\cdot\|$ represents the L_2 norm. Song documents with $K(d_i,d_j)$ close to one are very similar and can be viewed as very relevant in terms of attributes or "subjects."

Bottom-up clustering was performed on the database such that each song, d_i, in a cluster had a similarity of $K(d_i,d_j) > \tau_L$, where τ_L is a similarity threshold and d_j is the cluster mean. These clusters are based on the musical descriptions given by the Music Genome Project and do not contain non-acoustical features acting as "taxonomic noise." Unless stated otherwise, clusters which contain more than 10 songs were kept for later analysis.

3. DISCRIMINATIVE TRAINING REFINEMENT

The discriminative training (DT) technique described here provides smaller inter-cluster similarity, while increasing intra-cluster similarity Specifically, the columns of an MxK matrix, R, are updated in a training procedure to minimize misclassification, where K represents the number of clusters found in the previous section [13]. The training database is the set of songs in each of the retained clusters from Section 2.3 and each song is labelled with its corresponding cluster. This type of classifier is effective when the parametric form of the class distribution is unknown and when optimality in terms of estimating distributions is not necessarily equivalent to optimality in terms of classifier design [14].

The formulation starts by assigning the M-dimensional song vector x according to the misclassification function, $d_j(x,R)$:

$$\hat{j} = \arg\min_j d_j(x,R) = \arg\min_j[-g(x,R) + G_j(x.R)] \qquad (6)$$

where the jth column vector in R represents the jth cluster. The function $g_j(\cdot)$ is called the discriminate function and is often the dot product between x and the cluster mean of the jth cluster, r_j:

$$g_j(x,R) = r_j \cdot x = \sum_{i=1}^{M} r_{ji} x_i \qquad (7)$$

Likewise, the function $G_j(\cdot)$ is called the anti-discriminate function between x and the K-1 competing classes

$$G_j(x,R) = \left[\frac{1}{K-1}\sum_{k\neq j, 1\le k\le K} g_j(x,R)^{\eta}\right]^{1/\eta} \qquad (8)$$

where η is a positive number. As $\eta \to \infty$, the anti-discriminate function is dominated by the most competing class. The interpretation of (6)-(8) is to assign a vector to the class to which it is most similar when the scores from competing classes are also considered.

In order to find the optimal operating point given the training data, a smooth, differentiable 0-1 loss function is defined as

$$l_j(x,R) = \frac{1}{1+\exp(-\gamma d_j(x,R) + \theta)} \qquad (9)$$

where γ and θ are parameters that control the slope and shift of the sigmoid function, respectively. The overall empirical loss for the training database is

$$L(R) = \frac{1}{N_o}\sum_{i=1}^{N_0}\sum_{j=1}^{K} l_j(x_i,R)\mathbf{1}(x_i \in C_j) \qquad (10)$$

where N_o is the number of training samples from the clusters retained from Section 2.3 and $\mathbf{1}(\cdot)$ represents the indicator function, and C_j represents class i. Because $l_j(x,R)$ is a smooth, differentiable function, the overall empirical loss is minimized directly and the column vectors in R are updated using a gradient descent optimization procedure [13]. By combining the misclassification feature, d_j, into the objective function, the minimization of $L(R)$ increases the separation between classes while decreasing the distance between samples in a cluster. By using the discriminative training procedure, terms that strengthen intra-class similarity and increase inter-class distance are given more weight. Further, unlike LSA, term weights may obtain negative values, indicating that their inclusion in a test vector indicates it does not belong to the class in question.

4. EXPERIMENTAL RESULTS

The USPop2002 dataset [15] was used because of its popularity in MIREX Contests [1] and because many songs are described by the MGP. Of the 8764 tracks in the dataset, 3108 song descriptions were found. After text normalization, a list of 375 words was used in the analysis.

4.1. Song Clustering

SVD was performed on the word-document matrix following LSA. Singular values were kept based on the percentage volume of the full matrix. Specifically, if the singular values are ordered such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_Q$ then the minimum number, Q_o, is found such that

$$\frac{\sum_{i=1}^{Q_0} \lambda_i^2}{\sum_{i=1}^{Q} \lambda_i^2} \leq \tau_{vol} \qquad (11)$$

It was found that the number of clusters and the cluster sizes increase as τ_L decreases because the criteria for similarity is relaxed. The number of clusters did not change much as τ_{vol} was varied, which demonstrates that the attribute lists are largely noise-free. A possible reason is that (after text normalization), the final attribute set is fairly independent in meaning.

The larger clusters were examined when τ_{vol} = 0.80 and τ_L was varied. When τ_L = 0.95, the largest cluster contained six songs:

1. "Tiger" by Abba
2. "The Ballad of El Goodo" by Big Star
3. "Flowers" by New Radicals
4. "Simple Kind of Life" by No Doubt
5. "How's it Going to Be" by Third Eye Blind
6. "She Takes Her Clothes Off" by Stereophonics

While this might seem a bit eclectic, all songs had very similar attribute lists. Specifically, all songs were defined by a basic rock song structure, mixed acoustic and electric instrumentation, a vocal harmony, and major key tonality. In addition, all contain acoustic rhythm guitars except "Tiger," which contains an acoustic rhythm piano.

The other clusters in the τ_L = 0.95 were fairly small (3-4 songs) and usually contained a single artist (e.g., AC/DC, Nirvana, Blink 182). However, as these clusters grew, they started to obtain more variety in artists and individual artists also appeared in multiple categories. Clear cluster definitions also started to emerge, such as a group based on "basic rock structure," one with "electronica and disco influences", and a third containing "thin orchestration and radio friendly stylings."

4.2. Discriminative Trained Clusters

As stated in Section 2.4, discriminative training provides maximum separation in terms of misclassification by emphasizing features to strengthen within-class similarity, while pushing separate classes farther apart. Further, discriminative training on LSA unigram vectors has been shown to be as effective as trigram models [13]. To understand the difference in class separation between the unsupervised clusters versus more traditional taxonomies, the empirical risk was examined for both cases. The unsupervised clusters with more than 10 songs at thresholds of τ_{vol} = 0.85 and τ_L = 0.80 were fed to the DT classifier. In addition, a separate classifier was trained, but the labels for each song came from the genre label that could be found for the artist at allmusic.com.

Figure 2 shows the empirical loss as defined in (10). Convergence is quicker when using the unsupervised musical attribute clusters than clustering using genre labels. While genre does eventually reach a minimum of around 4% after 16000 iterations, the empirical loss is still higher than the unsupervised approach.

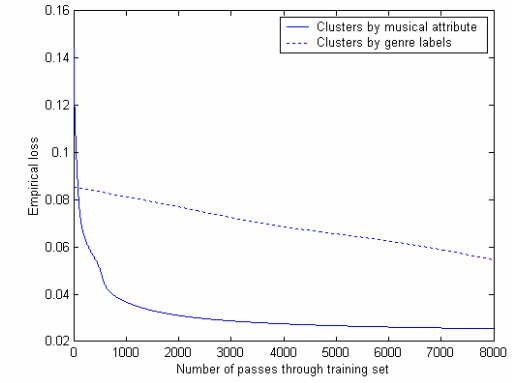

Figure 2. Empirical loss for the discriminatively trained song description clusters.

The drop in empirical loss from the discriminative training procedure is due to better separation between the different clusters. This can be seen from the change in the misclassification function described in (6), as is shown in **Error! Reference source not found.**. Specifically, the distribution is shifted left, indicating fewer misclassifications.

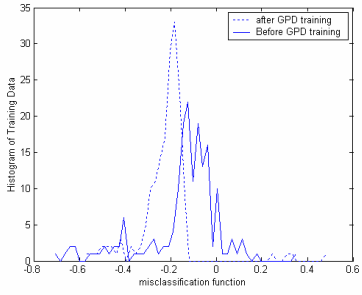

Figure 3. Histogram of the d-values in Eq. (6) before and after GPD training for the unsupervised clusters.

Another reason for the drop in empirical loss is that features which better represent a cluster are emphasized. Further, features that negatively correlate with a class will have a negative weight and other features, which do not increase performance, are given values close to zero. The example mean vector in Figure 4 demonstrates the effect of discriminative training on term weights. From a qualitative standpoint, the terms with the biggest positive and negative weight indicate this class contains many songs with acoustic guitars riffs and do not feature breathy (vocals), antiphony, or minor tonality.

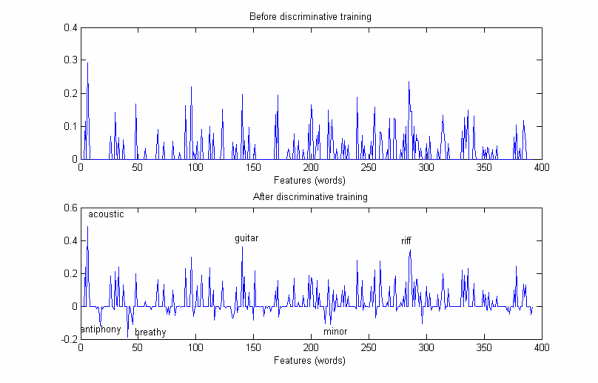

Figure 4. Attribute weights for an example cluster mean before and after discriminative training.

5. DESCRIPTORS FOR TRADITIONAL TAXONOMY

The previous section demonstrated that the unsupervised clusters performed better in providing acoustic-based descriptors than the traditional genre labels given in MIR applications. As online retailers obtain more of the market share for music distribution, one could expect for individual songs to obtain genre labels. However, cognitive studies show that even in terms of individual songs, humans often make genre decisions based on non-acoustical cues [5]. To test the impact of assigning genre labels at the song level, a classification experiment was performed using attributes from MGP.

Some attributes contained typical genre labels, such as "hip-hop roots" and "basic rock song structure." Songs with these attributes were manually labelled to the prescribed genre and only songs with a single genre attribute were used. A total of 25 genres were found and each genre was considered in a flat hierarchy. This was done to better understand the acoustic attributes that described potential sub-classes. Only genres with more than 40 songs were retained, resulting in the following genres: *rock*, *pop-rock*, *r&b*, *rap*, and *country*. LSA and discriminative training were performed on the retained songs. Further, bigram and unigram information was examined. In addition, a comparison was made between performing LSA prior to SVD and using simple word counts (i.e., no entropy-normalization was performed).

The empirical loss in **Figure 5** suggests that entropy normalization on unigrams slows convergence. Given the nature of musical attributes, this is not surprising. One difference between MIR and text information retrieval is that rarity does not necessarily translate to better retrieval results. For example, an important descriptor in human similarity judgements is *major* and *minor* tonality [5]. However, since almost every piece of Western music contains either/both *major* and *minor* tonality, the indexing power given to the terms is very low since it occurs in almost every song attribute list. Further, bigram counts appear to hurt overall accuracy. The most likely reason for this is that many bigrams are unimportant and better text normalization might improve performance.

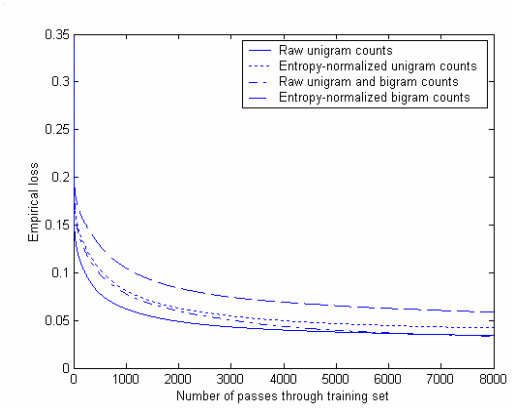

Figure 5. Empirical loss for unigram and bigram word counts and unigram and bigram entropy-normalized counts using MGP genre labels.

To test the generalization ability of the genre mean vectors found, another set of 5825 songs from MGP was classified. The resulting confusion matrix for the raw unigram counts is shown in Table 1.

The overall accuracy was found to be 80.26%, indicating that musical attributes do not consistently appear within a genre. These initial results indicate that even musicologists are unable to correlate acoustic descriptors on any temporal scale with genre taxonomies. If the gap between low-level cognitive features cannot describe the high-level attribute of genre association, there is little hope that low-level acoustic attributes will perform better. These results should not be interpreted as saying that genre recognition is an impossible task. Instead, this indicates that non-acoustical information may be necessary in order to accomplish the genre recognition task.

	C	PR	RB	RA	RK
C	49.65	2.8	2.45	0	45.1
PR	1.81	15.03	8.81	0	74.35
RB	0	14.35	38.12	0	47.54
RA	3.31	0.74	2.21	87.5	6.25
RK	1.34	4.72	1.68	0.09	92.16

Table 1. Confusion matrix for MGP genre labels. C = *country*, PR = *pop-rock*, RB = *r&b*, RA = *rap*, RK = *rock*

6. CONCLUSIONS

Cognitive studies have shown that genre and other similarity labels often ascribed to music contain some non-acoustical information, which is impossible for an acoustic classifier to learn. With the use of web-services, information on acoustic attributes can be gathered and taxonomy can be created. By seeking specific acoustic-based information, databases can be built such that ground-truth labels are given at the song level, based on measurable acoustic attributes, and exclude non-acoustic information. This reduces the potential of over-generalizing these systems to match an unrealizable performance level. Further, acoustic similarity can be defined in terms of relevance scores based on acoustic attributes.

This paper has presented an approach for designing acoustic MIR databases. First, musical similarity ground-truth labels were built by matching musical attributes using LSA and document clustering techniques. Further class separation was achieved using a discriminative training procedure, which not only finds terms that are important for the class in question, but also finds terms whose inclusion in a song indicates dis-similarity. Second, the variation within a genre and between different genres was examined using a similar procedure, but with ground-truth assigned by traditional genre labeling procedures. The results indicate that more information is used to assign genres than acoustic information.

7. REFERENCES

[1] "MIREX 2006 Results," [Web site] 2006, [2007 March 24], Available: http://www.music-ir.org/mirex2006/index.php/MIREX2006_Results.

[2] J.-J. Aucouturier and F. Pachet. "Representing musical genre: a state of the art," *J. New Music Research*, vol. 32, no. 1, pp. 1-12, 2003.

[3] H. G. Tekman and N. Hortacsu. "Aspects of stylistic knowledge: what are the different styles and why do we listen to them?" *Psychology of Music*, vol. 30, pp. 23-47, 2002.

[4] C. McKay and I. Fujinaga. "Musical genre classification: is it worth pursuing and how can it be improved?" *ISMIR 2006*, Victoria, Canada, pp. 101-107, 2006.

[5] D. Huron. *Sweet Anticipation: Music and the Psychology of Expectation*. MIT Press, 2006.

[6] B. Whitman and P. Smargdis. "Combining musical and cultural features for intelligent style detection," *ISMIR* 2002, pp. 47-52, 2002.

[7] J. Reed and C.-H. Lee, "A study on music genre classification based on universal acoustic models," *ISMIR 2006*, Victoria, Canada, pp. 89-94, 2006.

[8] A. Lamont and N. Dibben, "Motivic structure and the perception of similarity," *Music Perception*, vol. 18, no. 3, pp. 245-274, Spring 2001.

[9] J.H. Lee and J.S. Downie. "Survey of music information needs, uses, and seeking behaviors: preliminary findings," *ISMIR 2004*, Barcelona, Spain, 2004.

[10] C.D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[11] J. Bellegarda, "Exploiting latent semantic information in statistical language modelling," *Proc.* IEEE, vol. 88, no. 3, pp. 1279-1296, August 2000.

[12] G. Strang. *Linear Algebra and its Applications*. Harcourt, Inc., 1998.

[13] H.-K. Kuo and C.-H. Lee. "Discriminative training of natural language call routers," *IEEE Trans. Speech and Audio Proc.*, vol. 11, no. 1, pp. 24-35, Jan. 2003.

[14] B.-H. Juang, W. Chou, and C.-H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Trans. Speech and Audio Proc.*, vol. 5, no. 3, pp. 257-265, May 1997.

[15] A. Berenzweig, B. Logan, D. Ellis, and B. Whitman. "A large-scale evaluation of acoustic and subjective music-similarity measures." *Computer Music Journal*, vol. 28, iss 2, pp. 63-76, June 2004.