A Two-Stage Audio Retrieval Method for Searching Unannotated Audio Clips

Songhua Xu[1,2,3,*] Suchao Chen[1] Kevin Y. Yip[2] Francis C.M. Lau[3] Xueying Qin[1,4]

1: College of Computer Science and Technology, Zhejiang University
Hangzhou, Zhejiang, P.R. China, 310027
2: Department of Computer Science, Yale University
New Haven, Connecticut, USA, 06520-8285
3: Department of Computer Science, The University of Hong Kong
Pokfulam Road, Hong Kong, P.R. China
4: School of Computer Science and Technology, Shandong University
Jinan, Shandong, P.R.China, 250101
*: Correspondence author. Contact him by songhua DOT xu AT gmail DOT com.

Abstract

Traditional audio retrieval systems deal principally with audio clips having text descriptions. To retrieve unannotated audio clips is cumbersome because of the immaturity of content-based analysis and retrieval techniques. In this paper, we propose a two-stage audio retrieval method, consisting of a first stage of text-based retrieval and a second stage of content-based retrieval. This new retrieval method can be employed to retrieve audio clips from an audio collection having only partial text annotations, which is true of many online audio datasets. We have developed a prototype audio retrieval system based on our algorithm and carefully evaluated its performance. The results demonstrate the effectiveness of our new audio retrieval method. Our method can be generalized and applied to other kinds of non-textual data such as images and videos.

1. Introduction

Today's information retrieval techniques have achieved grand success in their application to text documents, which is testified to by the huge commercial profits generated by search engine companies such as Google and Yahoo!. In comparison, multimedia retrieval is at an infancy stage and no existing product or tool has offered a user satisfaction or popularity comparable to text-based search engines. In particular, retrieving audio clips which have no text annotations, an important function in many potential applications, is still largely an unsolved problem from the commerical point of view. This paper attempts to address this problem and proposes a new method for audio retrieval from audio collections with only partial textual annotations (i.e., only a subset of the clips have annotation).

Current audio retrieval systems rely heavily on the annotation texts when processing audio data [1]. These annotations include structured and unstructured meta data, e.g., the title, the singer and sometimes lyrics in the case of a song. Retrieving audio clips based on their associated texts can essentially be handled the same way as retrieving text documents. Unlike web pages, from which keywords can be automatically extracted by algorithms, extracting text annotation from audio files can be challenging and error-prone. Furthermore, in reality only a fraction of all audio files would have been manually annotated by users; and these annotations could be biased or too terse to be useful. Thus text-based audio retrieval schemes have limited applicability and reliability. In the worst case, to search for a clip with no text annotation calls for great patience and determination on the part of the user, even when the audio collection to be searched is reasonably small, say 200 clips.

Besides text-based retrieval, an alternative is content-based audio retrieval based on content similarity metrics. For example, there has been an active thread of work in querying music by humming or voice recognition [10,18] in which the input query is a short passage of music hummed by the user. The search engine would perform a content-based search by audio similarity. While these work did achieve some notable progress, dealing with audio clips without text annotation in general remains a difficult task because of the high dimensionality of the audio feature space as well as the fuzziness and subjectiveness of content similarity which depends very much on the user and

the query in question.

In this paper, we propose a two-stage audio retrieval method for retrieving audio clips from an audio collection with only partial text annotation. The first stage involves a text-based retrieval, which is followed by a content-based retrieval in the second stage. The key advantage of our method is its general applicability to many different types of media data. Due to space limitation, we focus on audio retrieval in this paper even though it should not be difficult to extend our idea to apply to other non-textual data, such as images and videos. We have conducted a series of audio retrieval experiments and obtained very positive results.

The rest of the paper is organized as follows. We first briefly survey the most related work on audio retrieval in Sec. 2. And then we introduce our two-stage audio retrieval method in Sec. 3. We report our experiment results in Sec. 4 and discuss the advantages and applicability of our method in Sec. 5. Finally, we conclude the paper in Sec. 6.

2 Related Work

The general problem of retrieving audio clips against a query has been a research topic for many years [8]; a number of audio indexing and retrieval technologies have been proposed, e.g., [6, 17, 25]. One key issue is the measurement of similarity between pairs of audio clips. A popular approach is to use the mel frequency cepstral coefficients or their variants to define the similarity between audio clips [15]. Some researchers have applied clustering techniques based on spectral features to form song signatures, which can then be used to compare different songs [16]. Acoustic and subjective music similarity have been proposed for large scale music retrieval [5]. Berenzweig et al. [4] employed Gaussian mixture models to represent objects such as songs in anchor spaces, and then used approximations of KL-divergence to define similarity measures to match human similarity labeling data. Aucouturier and Pachet [2] used Gaussian models of cepstrum coefficients to define timbral similarity for comparing music titles. Rhythmic [9] and sequential [6] information have also been used in defining music similarity. In addition to acoustic similarity, Barrington et al. [3] proposed a semantic similarity for retrieving audio data. Special purpose retrieval systems have also been developed for retrieving spoken documents [26], audio lectures [19], and news [24]. Most recently, Eck et al. [7] have studied how to automatically generate social tags for music which are untagged or poorly tagged to reduce the cold-start problem in recommender systems. Our work is also related to the problem of audio classification. Various standard machine learning techniques have been employed to address the problem, e.g., approaches based on support vector machines [11] and semi-supervised learning [14].

Our work is also related to research on deriving semantic descriptions for music retrieval [23]. Slaney [21] introduced a clustering method in a multi-dimensional vector space for semantic audio retrieval. Levy and Sandler [13] presented an algorithm for semantic space derivation for music from social tags. Sordo et al. [22] used music similarity to propagate text annotations for music. In comparison, our current work focuses on audio retrieval, but not the problem of annotation tag generation for music even though both problems are closely related.

There are a few existing audio retrieval systems [12]. For them, query by humming [10, 18] is a particularly popular research track. Most recently, Rice and Bailey [20] have proposed an audio file search system supporting both description text based search and sonic similarity based search. In their system, these two search options are separate and cannot be jointly used. Also, in contrast to our method proposed in this paper, they did not attempt to propagate text descriptions of audio files to other similar pieces.

3 Our Two-Stage Audio Retrieval Method

3.1 Notations

Let A_i's ($i = 1, \cdots, n$) be the set of all the candidate audio clips to be retrieved. Suppose there are n_a of them annotated with text descriptions. Without loss of generality, we assume they are $A_1, \cdots, A_{n_a}$. For each of them, we use $\mathbf{y_i}$ to represent the texts associated with A_i. The remaining $n_u = n - n_a$ audio clips are not annotated, i.e., $A_{n_a+1}, \cdots, A_n$. For each audio piece A_i, we derive a d-dimensional audio content feature vector $\mathbf{x}_i \triangleq (x_{i,1}, \cdots, x_{i,d})$. Each of the above audio features $x_{i,j}$ ($j = 1, \cdots, d$) is a real number within the range of 0 to 1. Currently, we consider 16 types of audio features, as listed in Table 1. Note that each of the first 6 features is a single number, but each of the latter 10 features is a time signal. For each of these latter 10 types of audio features, we compute its mean, standard deviation as well as the mean and standard deviation of the signal's first order forward finite difference. Thus each of these 10 feature signals produces 4 values. This results in an overall of $d = 6 + 10 \times 4 = 46$ features values. Finally, we normalize each of them to the range of 0 and 1 by dividing by the maximum corresponding feature value of an audio clip in our audio collection respectively.

3.2 First stage text-based audio retrieval

Given a user input query Q comprising one or a few keywords, we first conduct a text-based query using a conventional text-based retrieval method. Currently, we retrieve audio clips whose text annotations contain all the query

Table 1. The 16 audio features used in our experiments.

No.	Feature Name	No.	Feature Name
1	Rhythm Patterns	9	Total Loudness
2	Statistical Spectrum Descriptor	10	Mel Frequency Cepstrum Coefficient
3	Rhythm Histogram	11	Audio Spectrum Centroid
4	Auto-correlation	12	Audio Spectrum Rolloff
5	Log Attack Time	13	Audio Spectrum Spread
6	Temporal Centroid	14	Sone/Bark Bands
7	Audio Power	15	Zero-crossing Rate
8	Fundamental Frequency	16	Audio Spectrum Flatness

keywords, i.e., $\mathbf{y}_i \supseteq Q$. Since text-based retrieval is not the main focus of this paper, we skip the details, although we do plan to investigate and incorporate more advanced text query methods in the future. Let $\mathbf{r}^a_{n\times 1}$ be the query result vector where the i-th component of the result vector $\mathbf{r}^a(i) = 1$ if A_i is retrieved in the above text-based query process and $\mathbf{r}^a(i) = 0$ otherwise. Notice that the above vector $\mathbf{r}^a$ is a binary vector, i.e., each of its components $\mathbf{r}^a(i)$ is a binary value, indicating whether the audio clip A_i is retrieved under the query Q. Apparently, only those audio clips which have text annotations will be retrieved in the process. So for those audio clips which do not have text annotations, its corresponding $\mathbf{r}^a(i)$ is set to 0.

Our main focus is on the second stage, content-based retrieval. The goal of this stage is to construct another query result vector $\mathbf{r}^u_{1\times n}$ based on $\mathbf{r}^a$ which specifies whether an audio clip without text annotation is relevant to the given query Q. Different from $\mathbf{r}^a$, each component in $\mathbf{r}^u(i)$ is a real number between 0 and 1, indicating how likely a clip is relevant to the query. The higher the value is, the more likely the audio is related to the query. Given a certain threshold τ, those audio clips A_i's with their $\mathbf{r}^u(i) > \tau$ will be returned as the query result set.

3.3 Second stage content-based audio retrieval

The overall task of our second stage content-based audio retrieval is to construct $\mathbf{r}^u$ by identifying audio clips similar to the ones returned in the first stage text-based query process, the latter of which is indicated through the vector $\mathbf{r}^a$ as explained in Sec. 3.2. To do this, we first define a pairwise audio content similarity measure $s(\mathbf{A_i}, \mathbf{A_j})$ based on these two audio clips' audio feature vectors $\mathbf{x}_i$ and $\mathbf{x}_j$:

$$s(\mathbf{A}_i, \mathbf{A}_j) \triangleq \frac{1}{d}\sum_{k=1}^{d}\big(1 - (x_{i,k} - x_{j,k})^2\big). \qquad (1)$$

Recall $x_{i,k}$ and $x_{j,k}$ are the k-th components of the audio feature vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ respectively. Intuitively, two clips are more similar in content if they have closer feature values, and hence the above equation will give a larger similarity value. Here we simply weigh all the features equally. We will investigate more sophisticated techniques to optimally assign weights to balance multiple features in the future.

Once the above pairwise audio similarity measure is defined, we can derive an overall audio-to-query relevance score for the audio clip A_i as:

$$\mathbf{r}^u(i) \triangleq \max_{\mathbf{r}^a(j)=1} s(\mathbf{A}_i, \mathbf{A}_j). \qquad (2)$$

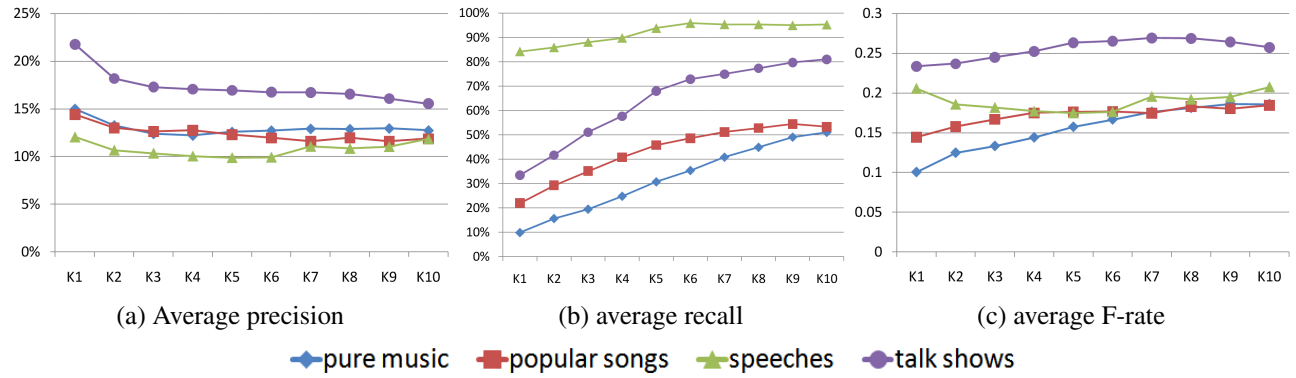
This equation computes the maximum similarity between $\mathbf{A}_i$ and an audio clip retrieved in the first stage of text-based querying. All the audio clips are then sorted in a descending order in terms of their corresponding $\mathbf{r}^u(i)$ values which are required to be larger than a user tunable threshold. Filtering can be applied so that an audio clip A_i is returned as a search result only if there are at least a certain number of clips retrieved in the first stage of text based querying process whose content similarity with A_i is above the threshold. To maximize the recall rate, in our current experiment, we do not use this filtering option. For applications in building audio recommender systems where precision is more emphasized than recall, users can turn on the option so that only the most confidently related audio clips will be returned. The above process essentially defines a similarity-based search result propagation process, which constitutes our second stage content-based audio retrieval.

4 Experiments

4.1 Data

We gathered 7335 clips of audio data from the Internet for our experiment, which are roughly classified into the four categories:

- *Pure Music*: We downloaded from the Internet 2147 audio clips of pure music. Each clip is annotated by the title of the song and the instrument name.
- *Popular Songs*: We acquired 3496 audio clips of popular songs from the Internet. Each clip is annotated by the title of the song, the singer's name, and the lyrics.
- *Public Speeches*: This dataset contains 234 clips of public speeches, downloaded from websites providing instructional materials for English as Second Language (ESL) studies. Each clip is associated with the full script of the speech as well as the speaker's name and the title of the speech.



(a) Average precision (b) average recall (c) average F-rate

Figure 1. Performance comparison between our basic method and the performance optimized method. The x-axis indicates the cases for $K_1, \cdots, K_{10}$ respectively, i.e. the situations when there are $1, \cdots, 10$ audio clips found in the first stage respectively. The y-axis shows the precision (a), recall (b), and F-rate (c) respectively.

- *Talk Shows*: This dataset contains 1458 audio clips of talk shows from an entertainment web site. Each talk show clip is annotated by the performer's name, the title of the performance and in some cases the script of the show.

4.2 Experiment Setup

In our experiments, we used an automatic procedure to evaluate the performance of our two-stage audio retrieval method. To generate testing queries, we adopt the following way to formulate queries for retrieval experimentation with different categories of audio data: 1) for pure music, we use the instrument name as the query keyword; 2) for popular songs, we use the singer's name; 3) for public speeches, we use the speaker's name; 4) for talk show programs, we use the performer's name. Our testing queries are constructed to look for all the audio clips in our database that are played by the instrument or by the given singer, or speaker or performer.

To evaluate the performance of our two-stage audio retrieval method, we notice the number of audio clips found in the first stage has a significant impact on the overall performance. We denote the situation when there are x audio clips found in the first stage as K_x. We report the performance evaluation data of our method for situations $K_1, \cdots, K_{10}$ respectively.

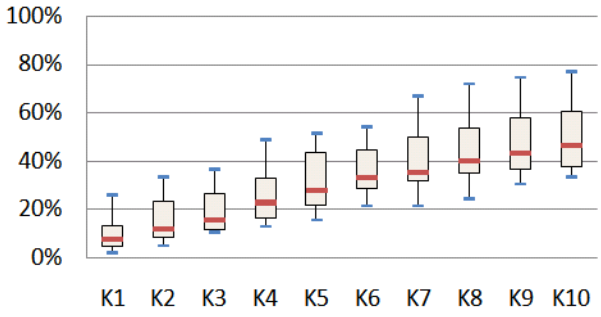
When evaluating the situation K_x, we randomly find x audio clips from our dataset whose text annotation contains the keyword in the query. We then take these clips as the result of our first stage text-based retrieval and hide the text annotation of all the other clips in our dataset. We then apply our method introduced in this paper to the dataset for audio retrieval. After that, we examine the text annotation originally associated with the audio clip to determine whether the retrieved one is relevant to the query by simply checking whether the annotation text contains the keyword in the query. To obtain the performance for each situation K_x, we repeat the above process five times and average the precision, recall, and F-rates. We report the average precision, recall and F-rate of all the experiments over the dataset of music, popular songs, speeches and talk shows respectively in Figure 1. This figure shows when more sample audio clips are found in the first stage text-based querying process, the precision of our method remains roughly the same while the recall rate becomes significantly better, which also results in improvement in the F-rate. Figure 2 employs boxplots (box-and-whisker diagrams) to examine the recall of our two-staged audio retrieval method for the four types of audio data respectively when different numbers of audio clips are found in the first stage text-based retrieval step. In each boxplot, we report the minimum, lower quartile (25^{th} percentile), median (50^{th} percentile), upper quartile (75^{th} percentile), and the maximum of the recall statistics of all the querying experiments performed over a certain type of audio data. As can be seen from the figure, the recall of our two-stage audio retrieval method is significantly improved.

5 Discussion

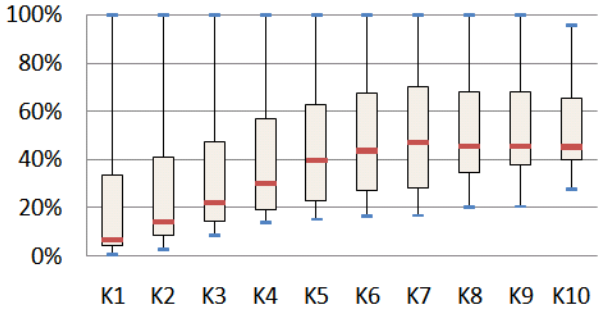
5.1 Advantages of our Two-stage Audio Retrieval Method

Our two-stage audio retrieval consists of a first stage text-based querying and a second stage content-based querying. There are three key advantages to our approach.

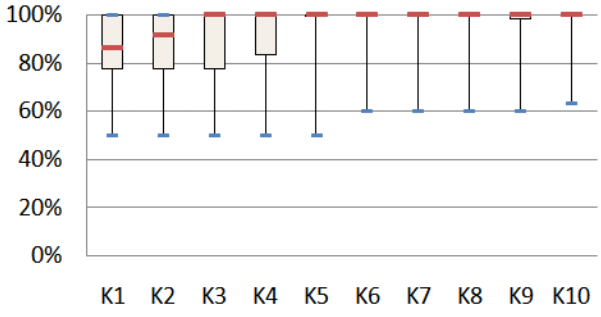
First, even though text-based searching is a well studied problem when compared with content-based audio retrieval, there are a large collection of audio clips online which are



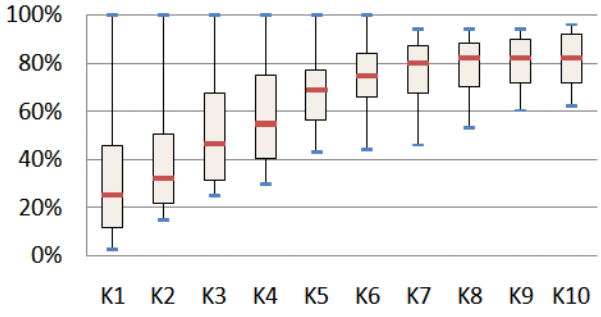
(a) Recall for pure music



(b) Recall for popular songs



(c) Recall for speeches



(d) Recall for talk shows

Figure 2. Boxplots over the recall of our two-stage audio retrieval method applied onto four datasets for the settings $K_1, \cdots, K_{10}$ respectively, i.e. the situations when there are $1, \cdots, 10$ audio clips found in the first stage.

either not annotated or poorly annotated. Text-based searching alone would lead to a low recall in most cases. We thus turn to the second stage content-based querying, which can significantly boost the recall rate.

Second, unlike text based queries, audio clips might be too clumsy to upload as query input by the ordinary users. In fact, in most cases, users probably will not have a suitable clip on hand to submit as a query. Using our two-stage query, there is no need for the user to provide a sample audio, which is much more convenient.

Third, our two rounds of querying can comprehensively consider multiple clues in inferring relevant audio clips. If a candidate audio clip is similar to two sample audio clips, it has a better chance to be relevant than the one which is similar to only one sample audio clip. Therefore, the more sample audio clips we have, the more likely and reliably the second stage content-based audio retrieval process can identify relevant audio clips. There are two essential parameters for the process: 1) the number of sample audio clips the candidate clip is relevant to; 2) how similar the candidate clip is to a sample clip. In this paper, we propose a two-stage audio retrieval method taking into consideration both parameters (Sec. 3.3).

5.2 Extending Our Algorithm to Other Non-textual Items

Our method is generically applicable to any content-based retrieval or recommender system in order to give recommendations on non-textual items. The basic procedure of our algorithm can be exactly applied. The only change needed is to replace the audio feature definition as presented in this paper by other domain specific features. For example, in the case of images, these domain specific features would be image features. In the case of videos, they would be video features.

6 Conclusion

In this paper, we have described a new audio retrieval method featuring a two-stage querying procedure. The procedure uses text data to obtain an initial set of audio clips which have associated textual annotation, and then propagates the annotation to the unannotated clips according to the content-similarity between multiple audio clips. Our new retrieval method enables retrieving audio files from an audio collection that lacks text annotation. Our experiment results show that our new method can perform satisfactorily over a range of audio files. In principle, our method is generally applicable to any content-based retrieval or recommender system to deal with non-textual items.

Acknowledgement

The research is partially supported by NSF of China (Grant No.60870003). This work has a patent pending.

References

[1] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749, 2005.

[2] J.-J. Aucouturier and F. Pachet. Music similarity measures: What's the use? In *Proceedings of the Third International Symposium on Music Information Retrieval*, pages 157–163, 2002.

[3] L. Barrington, A. Chan, D. Turnbull, and G. Lanckriet. Audio information retrieval using semantic similarity. In *ICASSP '07: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 725–728, April 2007.

[4] A. Berenzweig, D. P. W. Ellis, and S. Lawrence. Anchor space for classification and similarity measurement of music. In *ICME '03: Proceedings of the 2003 International Conference on Multimedia and Expo*, pages 29–32, Washington, DC, USA, 2003. IEEE Computer Society.

[5] A. Berenzweig, B. Logan, D. P. W. Ellis, and B. P. W. Whitman. A large-scale evaluation of acoustic and subjective music-similarity measures. *Computer Music Journal*, 28(2):63–76, 2004.

[6] M. Casey and M. Slaney. The importance of sequences in musical similarity. *ICASSP '06: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 5:5–8, May 2006.

[7] D. Eck, P. Lamere, T. Bertin-Mahieux, and S. Green. Automatic generation of social tags for music recommendation. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *NIPS '08: Proceedings of Advances in Neural Information Processing Systems*, pages 385–392. MIT Press, Cambridge, MA, 2008.

[8] J. Foote. An overivew of audio information retrieval. *Multimedia Systems*, 7(1):2–10, 1999.

[9] J. Foote, M. Cooper, and U. Nam. Audio retrieval by rhythmic similarity. In *Proceedings of the Third International Conference on Music Information Retrieval*, pages 265–266, 2002.

[10] A. Ghias, J. Logan, D. Chamberlin, and B. C. Smith. Query by humming: musical information retrieval in an audio database. In *MULTIMEDIA '95: Proceedings of the Third ACM International Conference on Multimedia*, pages 231–236, New York, NY, USA, 1995. ACM.

[11] G. Guo and S. Li. Content-based audio classification and retrieval by support vector machines. *IEEE Transactions on Neural Networks*, 14(1):209–215, Jan 2003.

[12] K. Hoashi, H. Ishizaki, K. Matsumoto, and F. Sugaya. Content-based music retrieval using query integration for users with diverse preferences. In *Proceedings of International Conference on Music Information Retrieval*, pages 463–466, 2007.

[13] M. Levy and M. Sandler. A semantic space for music derived from social tags. In *Proceedings of International Conference on Music Information Retrieval*, pages 411–416, 2007.

[14] T. Li and M. Ogihara. Music artist style identification by semi-supervised learning from both lyrics and content. In *MULTIMEDIA '04: Proceedings of the 12th Annual ACM International Conference on Multimedia*, pages 364–367, New York, NY, USA, 2004. ACM.

[15] B. Logan. Mel frequency cepstral coefficients for music modeling. In *Proceedings of the First International Symposium on Music Informaiton Retrieval*, 2000.

[16] B. Logan and A. Salomon. A music similarity function based on signal analysis. In *ICME '01: Proceedings of IEEE International Conference on Multimedia and Expo*, pages 745–748, 2001.

[17] J. Makhoul, F. Kubala, T. Leek, D. Liu, L. Nguyen, R. Schwartz, and A. Srivastava. Speech and language technologies for audio indexing and retrieval. *Proceedings of the IEEE*, 88:1338–1353, 2000.

[18] B. Pardo and W. P. Birmingham. Query by humming: How good can it get? In *Proceedings of Workshop on Music Information Retrieval*, pages 107–109, 2003.

[19] A. Park, T. J. Hazen, and J. R. Glass. Automatic processing of audio lectures for information retrieval: Vocabulary selection and language modeling. volume 1, pages 497–500, 2005.

[20] S. V. Rice and S. M. Bailey. A system for searching sound palettes. In *Eleventh Biennial Symposium on Arts and Technology*, 2008.

[21] M. Slaney. Semantic-audio retrieval. In *ICASSP '02: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 4108–4111, 2002.

[22] M. Sordo, C. Laurier, and O. Celma. Annotating music collections:how content-based similarity helps to propagate labels. In *Proceedings of International Conference on Music Information Retrieval*, pages 531–534, 2007.

[23] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet. Towards musical query-by-semantic-description using the cal500 data set. In *SIGIR '07: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 439–446, New York, NY, USA, 2007. ACM.

[24] G. Tzanetakis and M.-Y. Chen. Building audio classifiers for broadcast news retrieval. In *Proceedings of IEEE Workshop on Image Analysis for Multimedia Interactive Services*, 2004.

[25] G. Tzanetakis and P. Cook. Audio information retrieval (AIR) tools. In *Proceedings of International Symposium on Music Information Retrieval*, 2000.

[26] B. Zhou and J. H. L. Hansen. SpeechFind: An experimental on-line spoken document retrieval system for historical audio archives. In *Proceedings of International Conference on Spoken Language Processing*, volume 3, pages 1969–1972, 2002.