Audio Similarity Measure by Graph Modeling and Matching

Yuxin Peng[1], Chong-Wah Ngo[2], Cuihua Fang[1], Xiaoou Chen[1], and Jianguo Xiao[1]
[1]Institute of Computer Science and Technology, Peking University, Beijing 100871, China
pengyuxin@icst.pku.edu.cn
[2]Department of Computer Science, City University of Hong Kong, Hong Kong, China
cwngo@cs.cityu.edu.hk

ABSTRACT

This paper proposes a new approach for the similarity measure and ranking of audio clips by graph modeling and matching. Instead of using frame-based or salient-based features to measure the acoustical similarity of audio clips, segment-based similarity is proposed. The novelty of our approach lies in two aspects: segment-based representation, and the similarity measure and ranking based on four kinds of similarity factors. In segment-based representation, segments not only capture the change property of audio clip, but also keep and present the change relation and temporal order of audio features. In the similarity measure and ranking, four kinds of similarity factors: acoustical, granularity, temporal order and interference are progressively and jointly measured by optimal matching and dynamic programming, which guarantee the comprehensive and sufficient similarity measure between two audio clips. The experimental result shows that the proposed approach is better than some existing methods in terms of retrieval and ranking capabilities.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval Models.

General Terms

Algorithms, Experimentation, Measurement, Theory.

Keywords

Audio similarity measure, audio retrieval.

1. INTRODUCTION

With the drastic advances of the audio and music content on the internet, there is an increase in the demand for audio content analysis, retrieval and summarization. In these techniques, content-based similarity measure is a critical fundamental step. In this paper, we propose a new approach for the similarity measure and ranking of audio clips by graph modeling and matching.

Existing approaches on audio clip retrieval can be classified into two categories: frame-based features [1][2] and salient-based features [3-5]. In frame-based features methods [1][2], a long audio clip is divided into many frames to catch the short time property. The features are extracted from each frame and their mean and standard deviation are calculated to form the feature vector of the audio clip. However, the frame features in an audio clip usually vary greatly along the time line. The mean and standard deviation of frames cannot give an accurate presentation of such property in audio clips. To complement the drawback, the methods in [3-5] propose to extract the salient characteristics or dominant features to present the change property of audio clips. In [3, 4], structure pattern is proposed for the similarity measure of audio clips, which describes the structural characteristics of both temporal and spectral features. In [5], dominant feature vectors are extracted from audio clips to represent the multiple salient characteristics of the clip. The methods in [3-5] are reasonable in capturing the change property of audio clip along the time line. Nevertheless, the salient characteristics are based on the statistical features of an audio clip, which cannot keep and present the change relation and order of audio clip along time line.

In audio retrieval task, frame-based representation, in general, is intuitive because audio frame is the basic structure in audio. However, due to the excessive number of frames in an audio clip, the mean and standard deviation of frames are then utilized to represent the audio clip for data reduction purpose, which is too rough to describe and represent the content change of audio clip. Similar to frame-based, the salient-based representation cannot solve this problem. In addition, *shot*, as the basic structure composed of frames in video domain, has been proved to be effective for the similarity measure of video clips [7]. In this way, a video clip is composed of shots, while a shot is composed of video frames. Motivated by the idea, we exploit a structural representation, namely *audio segment*, for the similarity measure of audio clips. Similar with shot in video domain, audio segment is a series of audio frames that are acoustically homogeneous, and the clip characteristics are represented by segment-based features. In an audio clip, the number of segment is only decided by its content change and has nothing with its duration. Since audio segment is acoustically homogeneous in general, the segment-based representation can capture the change property of audio clip. In addition, it keeps and presents the change relation and order of audio features because an audio clip is divided into physical segments along the time line. Furthermore, audio segment is semantically richer than audio frame, which is more appropriate for the similarity measure of audio clips.

Suppose audio clips are divided into several segments, the next problem will be how to measure the audio clip similarity according to the segment-based representation. Because an audio clip is divided into few segments, the similarity of two clips can be measured by their segment similarity. Then the similarity measure between two clips can be modeled as a weighted bipartite graph: every vertex in a bipartite graph represents one segment in an audio clip, and the weight of an edge represents the similarity value for a pair of segments between two audio clips. The bipartite graph simulates the many-to-many mapping among segments between two clips, as shown in the left graph of Figure 1. Certain criterion is demanded to measure the similarity based on the bipartite graph representation. An intuitive idea will be using one-

MM'06, October 23–27, 2006, Santa Barbara, California, USA.
ACM 1-59593-447-2/06/0010.

to-one matching among segments between two clips to measure the clip similarity. This is because every segment, as a part of an audio clip, will have its own effect for the final clip similarity. One-to-one matching could guarantee that every segment in an audio clip can be matched to the similar segment in another clip only once, while one-to-many matching will compute repeatedly the similarity among one segment with many segments, which will amplify the effect of one segment similarity for the final similarity measure. So, we employ a one-to-one matching algorithm, namely optimal matching (OM), to compute the maximum weight of the bipartite graph as the acoustical similarity value under the one-to-one mapping constraint, as shown in the right graph of Figure 1.

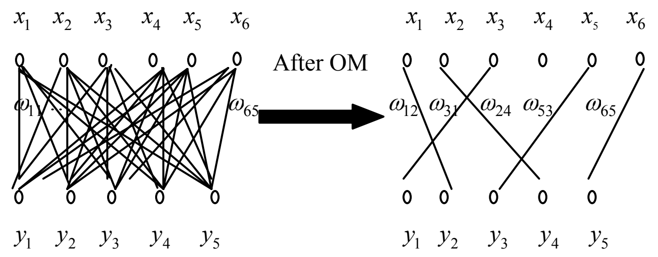
x_1 x_2 x_3 x_4 x_5 x_6
After OM
ω_1
ω_{65}
x_1 x_2 x_3 x_4 x_5 x_6
ω_{12} ω_{31} ω_{24} ω_{53} ω_{65}
y_1 y_2 y_3 y_4 y_5
y_1 y_2 y_3 y_4 y_5

Fig. 1. Optimal Matching

After one-to-one mapping, some segments cannot be matched between two audio clips, as shown the vertex x_4 in the right graph of Figure 1. In addition, the temporal order similarity between two clips should be also taken into consideration. Both factors will affect the similarity ranking of audio clips. In our approach, four kinds of similarity factors are measured: acoustical, granularity, temporal order, and interference factors. Granularity models the degree of one-to-one segment matching between two clips, order factor measure the temporal order similarity between two clips, while interference models the percentage of unmatched segments. In our approach, acoustical and granularity are measured by OM, temporal order similarity is evaluated by dynamic programming (DP). The measure of interference is based on the output of OM.

2. AUDIO PREPROCESSING

Preprocessing includes audio segmentation, feature representation and segment similarity measure. Bayesian Information Criterion (BIC) in [6] is employed to locally detect the single changes in the audio clip within a sliding window of variable size. Basically an audio clip is divided into several segments by the detected change points, and every segment is a series of audio frames that are acoustically homogeneous. In this way, the frames in a segment have nearly no change, so we use the mean of the feature values in all frames of the segment to represent every audio segment. Let the feature vector of an audio segment s_i be $\{f_{i1}, f_{i2}, \ldots\}$, the similarity between two audio segments s_i and s_j is defined as

$$Sim(s_i, s_j) = \exp(-Dis\tan ce(s_i, s_j)/2) \qquad (1)$$

$$Dis\tan ce(s_i, s_j) = \left(\sum_{p=1}^{n} (f_{ip} - f_{jp})^2 \right)^{\frac{1}{2}}$$

(2)

The distance function $Dis\tan ce(s_i, s_j)$ is Euclidean Distance of the feature vectors between two segments s_i and s_j. We utilize Eqn (1) to normalize $Dis\tan ce(s_i, s_j)$ to [0,1]. In the proposed approach, two types of features are computed for each audio frame: (1) log energy; (2) 12 order Mel-Frequency Cepstral Coefficients (MFCC). In this way, an audio frame is represented by a 13-dimensional feature vector. The first dimension feature is log energy, and the others are represented by MFCC features. Thus, a segment is also represented by a 13-dimensional vector.

3. CLIP-BASED SIMILARITY MEASURE

3.1 Optimal Matching (OM)

In our previous work, optimal Matching (OM) has been employed for the similarity measure of video clips [7]. In this section, we will extend OM to measure the similarity of audio clips. A weighted bipartite graph is constructed to model the similarity between two audio clips, and then OM is employed to compute the maximum weight of the bipartite graph as the acoustical similarity value under the one-to-one mapping constraint. Given two audio clips X and Y_k, a weight bipartite graph G_k is constructed as follows:

- Let $X = \{x_1, x_2, \ldots, x_p\}$ as a query clip with p audio segments and x_i represents an audio segment in X.
- Let $Y_k = \{y_1, y_2, \ldots, y_q\}$ as the k^{th} clip with q audio segments in an audio database Y and y_j is an audio segment in Y_k.
- Let $G_k = \{X, Y_k, E_k\}$ as a weighted bipartite graph constructed by X and Y_k. $V = X \cup Y_k$ is the vertex set while $E_k = (\omega_{ij})$ is the edge set. ω_{ij} represents the segment-based similarity value between x_i and y_j.

The constructed bipartite graph G_k is a complete weighted graph, which means every vertex x_i in X has an edge with every vertex y_j in Y_k and vice versa. That is to say, G_k has $p \times q$ edges. To measure effectively the similarity between two clips, OM is employed to maximize the total weights of matching under the one-to-one mapping constraint. The output of OM is a weighted bipartite graph G_{OM} where one segment in X can match with at most one segment in Y_k and vice versa. The similarity of X and Y_k is assessed based on the total weight in G_{OM} as follows

$$Sim_{OM}(X, Y_k) = \frac{\sum \omega_{ij}}{\max(p, q)}$$

(3) where the similarity is normalized by the maximum value of p and q, which are the number of audio segments in the query clip X and the clip Y_k relatively. The implementation of OM is based on Kuhn-Munkres algorithm [8]. The running time of OM is $O(n^4)$ where $n = p + q$ is the total number of vertices in G_k.

3.2 Dynamic Programming (DP)

Given a bipartite graph G_{OM} computed by OM, the similarity of two clips based on the temporal order of audio segments matching can be formulated by DP. Denote C as a cost matrix indicating the number of segments pairs that are matched along the temporal order, we have

$$c[i,j] = \begin{cases} 0 & i = 0 \text{, or } j = 0 \\ c[i-1, j-1] + 1 & i, j > 0 \text{ , } (x_i, y_j) \in M \\ \max(c[i, j-1], c[i-1, j]) & i, j > 0 \text{ , } (x_i, y_j) \notin M \end{cases}$$

(4)

where M is the optimal matching that contains the set of matched pairs formed by OM. The running time of Eqn (4) is $O(pq)$, where p and q are, respectively, the number of segments in X and Y_k. The similarity between two clips based on the temporal order is defined as

$$Sin_{DP}(X, Y_k) = \frac{C[p,q]}{p}$$

(5)

3.3 Interference factor (IF)

The IF counts the number of unmatched segments in G_{OM} . *i.e.*, $p + q - 2 \times |M|$. The similarity between two clips based on IF is

$$Sim_{IF}(X, Y_k) = \frac{2 \times |M|}{p + q}$$

(6)

Since the values of $|M|$, p and q are known, $Sim_{IF}(X, Y_k)$ can be computed in $O(1)$ time.

3.4 Clip Similarity

Given X and Y_k, the similarity is measured jointly by the degree of acoustical and granularity similarity, the temporal order of matching, and interference factor as follows:

$$Sim_{clip}(X, Y_k) = \sum_{i \in \{OM, DP, IF\}} \alpha_i Sim_i(X, Y_k) \tag{7}$$

where $\sum_i \alpha_i = 1$ are the weights of different similarity. The value of α_i controls the ranking of similar clips. In the similarity measure of audio clips, the degree of acoustical and granularity similarity, which is computed by OM, and reflect the proximity and number of matching segments respectively, is more effective than temporal order (DP) and interference factor (IF). Thus, we set $\alpha_{OM} > \alpha_{DP} = \alpha_{IF} (\alpha_{OM} = 0.4, \alpha_{DP} = \alpha_{IF} = 0.3)$ in experiments.

4. EXPERIMENTS

To evaluate the performance of the proposed approach, we set up a database with 1000 audio clips, which includes the database of Muscle Fish. The Muscle Fish database is extensively employed for the experimental evaluation of audio clip retrieval [3-5], which includes many kinds of sounds, such as *animals, human, vehicles, machines, music, weapon* and so on. In addition, the experimental database also includes some commercial clips. In total, the average time of every audio clip in the experimental database is 9 seconds.

In the database, all audio streams are down-sampled into 44k Hz and mono-channel. Each frame is 512 samples (23ms) with 25% overlapping. In the 1000 audio clips, 500 clips have the relevant clips, while the other 500 clips only appear one time in the database. The relevant clips, although belong to the same kind of sound, have different sound property. Overall, all the 500 clips with one or more relevant clips are selected as the query clips for a comprehensive performance comparison. Four methods are experimented for comparison, including the proposed approach, Gu's approach [5], L_2 distance, and Kullback-leibler distance [10]. In the above methods, the frame features are represented by 13-dimensional feature vector, the first dimension feature is log energy, and the others are represented by MFCC features, as the same with our approach in Section 2 for objective comparison. The major differences among the four approaches are summarized in Table 1.

Table 1. Comparison among our approach and other three methods

	Our approach	Gu's approach[5]	K-L distance[10]	L_2 distance
representation	segment-based	dominant-based	frame-based	frame-based
Similarity	four factors	acoustical	acoustical	acoustical
Measure	OM, DP	dominant feature	K-L distance	L_2 distance

4.1 Clip Retrieval

Recall and precision are adopted to evaluate the retrieval performance of audio clips. The recall and precision for four methods are shown in Figure 2 and Figure 3. The proposed approach outperforms other three methods in term of recall and precision, while Gu's approach, K-L distance and L_2 distance achieve almost same recall and precision. By manually investigating the retrieval results, we find the advantage of our approach is mainly due to: (1) Segment-based features can effectively represent the audio clips, which guarantee the effectiveness of clip-based similarity measure. (2) OM provides an effective mechanism for the similarity measure among audio segments by one-to-one matching.

4.2 Clip Ranking

In this experiment, our aim is to compare the ranking capability of these approaches. AR (average recall) and ANMRR (average normalized modified retrieval rank) are adopted to evaluate audio clip ranking performance [9]. The values of AR and ANMRR range from [0, 1]. A *high* value of AR denotes the superior ability in retrieving relevant clips, while a *low* value of ANMRR indicates the high retrieval rate with relevant clips ranked at the top [9]. Experimental results on AR and ANMRR for four methods are shown in Table 2. The proposed approach outperforms other three methods in term of AR and ANMRR. By tracing the details of experimental results, we found the acoustical similarity measure of frame-based or dominant-based features in other three methods cannot always give satisfactory results. In contrast, the proposed similarity measure of four factors based on OM and DP, can rank most of relevant clips at the top-*k* ranked list (*k* depends on the number of relevant clips [9]).

Currently, on a Pentium-4 2.4GHz machine with 512M memory, the average retrieval time for a query by our approach is approximately 0.258 second. By investigation, we found the average number of segments is 10 in our database, although many of them have a long duration. Therefore, although OM is not a linear time algorithm, it is still efficient even in a large database, since most of audio clips are divided into few segments according to their content change. This implies that the bipartite graphs constructed by audio clips have less vertices, which lead our approach to a faster retrieval speed.

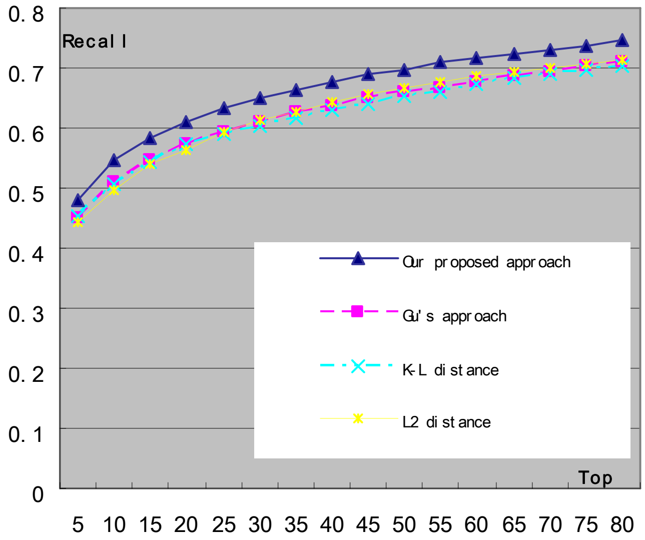

Fig. 2. Recall comparison of the four methods.

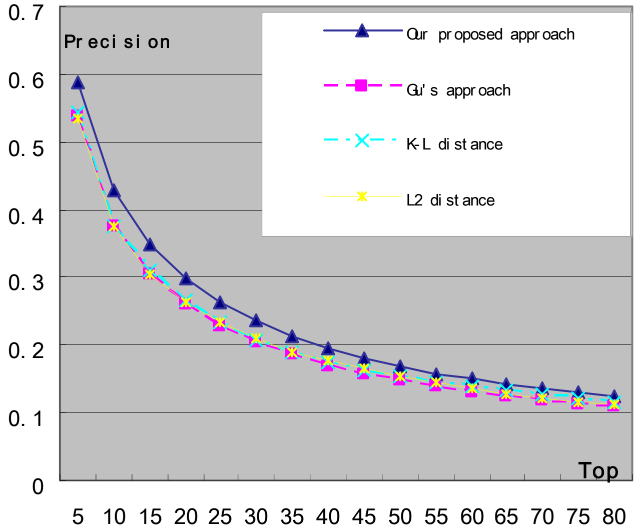

Fig. 3. Precision comparison of the four methods.

Table 2. AR and ANMRR for performance comparison of four methods

	Our Approach	Gu's Approach[8]	K-L Distance[13]	L_2 Distance
AR	0.72	0.66	0.67	0.66
ANMRR	0.26	0.33	0.32	0.33

5. CONCLUSIONS

We have presented a new approach for the similarity measure and ranking of audio clip based on graph modeling and matching. Four kinds of similarity factors: acoustical, granularity, temporal order and interference are progressively and jointly measured by optimal matching and dynamic programming, depending on the nature of the similarity factors. The experimental results show the effectiveness of our proposed approach.

In the future, more experiments will be conducted and analyzed. On one hand, we are interested in computing the optimal weight of OM, DP and IF based on the individual experiment analysis. On the other hand, we strive to investigate the more effective yet efficient matching methods based on segment representation for audio clip retrieval.

6. ACKNOWLEDGMENT

The work described in this paper was fully supported by the National Natural Science Foundation of China under Grant No. 60503062.

7. REFERENCES

[1] L. Lu, H. J. Zhang and H. Jiang, "Content Analysis for Audio Classification and Segmentation", *IEEE Transactions on Speech and Audio Processing*, Vol.10, No.7, pp.504-516, Oct. 2002.

[2] E. Wold, T. Blum, D. Keislar, and J. Wheaton, "Content-based Classification, Search, and Retrieval of Audio", *IEEE Multimedia*, 3(3): 27-36, 1996.

[3] R. Cai, L. Lu, H. J. Zhang and L. H. Cai, "Using Structure Patterns of Temporal and Spectral Feature in Audio Similarity Measure", *ACM Multimedia Conference*, pp.219-222, Berkeley, CA, Nov. 2-8, 2003.

[4] R. Cai, L. Lu, H. J. Zhang and L. H. Cai, "Improve Audio Representation by Using Feature Structure Patterns", *IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP)*, Vol IV, pp. 345-348, Montreal, Canada, May 17-21, 2004.

[5] J. Gu, L. Lu, R. Cai, H. J. Zhang and J. Yang, "Dominant Feature Vectors Based Audio Similarity Measure", *Pacific-Rim Conference on Multimedia(PCM)*, 2, pp.890-897, Nov 30-Dec 3, Tokyo, Japan, 2004.

[6] M. Cettolo and M. Vescovi, "Efficient Audio Segmentation Algorithms based on the BIC", *ICASSP*, *2003*.

[7] Y. Peng, and C. W. Ngo, "Clip-Based Similarity Measure for Query-Dependent Clip Retrieval and Video Summarization", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 16, No. 5, pp. 612-627, May 2006.

[8] W. S. Xiao, "*Graph Theory and Its Algorithms*", Beijing Aviation Industrial Press, 1993.

[9] MPEG video group, "Description of Core Experiments for MPEG-7 Color/Texture Descriptors", ISO/MPEGJTC1/ SC29/WG11 MPEG98/M2819, July, 1999.

[10] Z. Liu and Q. Huang, "Content-based Indexing and Retrieval-by-Example in Audio", *IEEE International Conference on Multimedia and Expo(ICME)*, 2000.