

Dominant Feature Vectors Based Audio Similarity Measure

Jing Gu^{*1}, Lie Lu², Rui Cai³, Hong-Jiang Zhang², and Jian Yang¹

¹ Dept. of Electronic Engineering, Tsinghua Univ., Beijing, 100084, China

² Microsoft Research Asia, Beijing, 100080, China

³ Dept. of Computer Science and Technology, Tsinghua Univ., Beijing, 100084, China

Abstract. This paper presents an approach to extracting dominant feature vectors from an individual audio clip and then proposes a new similarity measure based on the dominant feature vectors. Instead of using the mean and standard deviation of frame features in most conventional methods, the most salient characteristics of an audio clip are represented in the form of several dominant feature vectors. These dominant feature vectors give a better description of the fundamental properties of an audio clip, especially when frame features change a lot along the time line. Evaluations on a content-based audio retrieval system indicate an obvious improvement after using the proposed similarity measure, compared with some other conventional methods.

1 Introduction

A fundamental step of the audio content analysis is similarity measure based on the various features. In a general way, one extracts several features, including temporal and spectral features, from an audio clip (usually last for several seconds), then, the similarity measure between two audio clips is based on the feature vectors constructed by those features. In most cases, the clip is too long, so that it is usually divided into several 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 final feature vector of the audio clip. Such clip-based statistical features have proved their effectiveness in many previous works such as [1][2]. However, the frame features of an audio clip usually change much along the time line, that is, there usually exist more than one salient characteristic in the clip. Only the mean and standard deviation of frame features can not give an accurate presentation of the property of such audio clips [3].

For example, Fig. 1 (a) illustrates the spectrogram of a sound of applause. The sound shows periodicity and there are approximately two salient characteristics in the clip, one is for sound period and the other for silence period. Fig. 1 (b) gives another example of a sound made by a jet plane flying over the heads. The spectrogram shows an obvious spectral change due to the “Doppler

* This work was performed when the first author was a visiting student in Media Computing Group, Microsoft Research Asia

effect”. Therefore, the characteristics are quite different between two halves of the sound clip. Thus, two or even more salient characteristics are needed to describe this sound. In both cases, neither the mean nor the standard deviation of the frame features gives accurate description of different salient characteristics. Much information will be lost if one just uses them and thus will lead to inaccurate similarity measure in some cases. Therefore, it would be better to find a way to directly represent the most distinct characteristics of an audio clip.

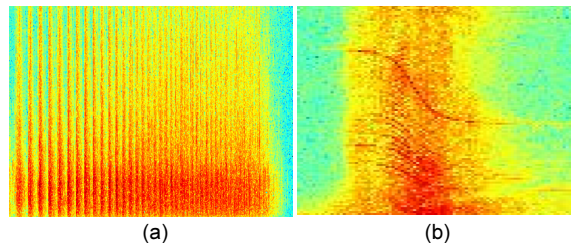


Fig. 1. Sound spectrograms of (a) *applause* and (b) *jet plane*.

In this paper, we propose a new approach to obtain salient characteristics of an audio clip by employing subspace decomposition on the set of frame-based features. The obtained dominant feature vectors describe the most salient characteristics of an audio clip, and can represent the audio clip better than the mean and standard deviation. The number of the dominant feature vectors needed to describe an audio clip is related to the feature variation in the corresponding clip. For example, it needs more dominant feature vectors if the characteristics of the clip change significantly; and less are needed when the characteristics keep stationary.

Suppose an audio clip has several salient characteristics or several dominant feature vectors, it is intuitive that further analysis should be implemented based on these dominant feature vectors. According to this, we further propose a new similarity measure between two audio clips. The proposed similarity measure considers the similarity between each pair of dominant feature vectors, and thus keeps the richness of sound property but reduces the noise.

The rest of the paper is organized as follows. Section 2 presents the detail approach to obtain the dominant feature vectors from an audio clip. Section 3 proposes a new similarity measure between clips and compares it with other conventional similarity measures. The evaluation of the proposed approach is given in the Section 4.

2 Dominant Feature Vectors

In this section, we extract the dominant feature vectors from an audio clip, in order to represent the multiple salient characteristics of the clip.

Assuming that an audio clip is divided into N frames, from each of which an n -dimensional feature vector is extracted and normalized to be zero mean and unit variance over the whole database. The normalized time-varying feature vectors of a clip can be represented by $X = (x_1, x_2, \dots, x_N)$, which is an n -by- N matrix and x_i ($i = 1, 2, \dots, N$) is the feature vector of the i^{th} frame. Now we want to obtain several dominant feature vectors which may give a good description of the audio clip, especially when several salient characteristics exist. Fortunately, we can achieve this object by employing eigen-decomposition on the covariance matrix of the frame based feature vector.

The n -by- n covariance matrix can be estimated as following:

$$C = \frac{1}{N} X X^T \quad (1)$$

By eigen-decomposition, the covariance matrix is decomposed as:

$$C = Q^T \Lambda Q = \sum_{i=1}^n \lambda_i q_i q_i^T \quad (2)$$

where $Q = (q_1, q_2, \dots, q_n)$ is an orthogonal matrix, q_i ($i = 1, 2, \dots, n$) is the eigenvector of C , and $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$ is a diagonal matrix of non-negative real eigen-values ordered such that $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$.

The main idea of this eigen-decomposition is to find the vectors which best describe the characteristics of the set of frame feature vectors, within the space spanned by them. Generally, the eigen-vectors associated with large eigen-values represent the dominant information and can be considered as dominant feature vectors, while those eigen-vectors with small eigen-values have little contribution and can be considered as introduced by noise. Therefore the eigen-vectors associated with large eigen-values are the dominant feature vectors we want to obtain. The corresponding eigen-values can be considered as the importance or the contribution of the dominant feature vectors.

The number of dominant feature vectors needed to represent an audio clip is related to the characteristic variation of a clip. It needs more dominant feature vectors if the characteristics of the clip change much. Considering the eigen-values represent the contribution of the corresponding eigen-vectors, a general way of choosing the number of dominant feature vectors is as follows:

$$m = \arg \min_k \left\{ \sum_{i=1}^k \lambda_i / \sum_{i=1}^n \lambda_i \geq \eta \right\} \quad (3)$$

where m is the number and the threshold $\eta \in (0, 1)$.

The m dominant feature vectors actually span an m -dimensional subspace which is the best approximation of the original n -dimensional eigen-space, suppressing the effect of noise. We call the m -dimensional subspace “*signal subspace*”. The remaining $(n - m)$ -dimensional subspace is called “*noise subspace*”. Correspondingly, the noise-reduced covariance matrix can be represented by

$$C = \sum_{i=1}^m \lambda_i q_i q_i^T \quad (4)$$

It should be noted that our approach to dominant feature vector extraction is totally different with traditional PCA applications. PCA is traditionally used to remove the noisy feature dimensions. However, our approach is used to remove the noisy feature vectors so that the dimension of each feature vector is not decreased. Moreover, dominant feature vectors is performed on an individual audio clip and form a "signal subspace" which represents the most salient characteristics of the corresponding audio clip, while PCA usually is based on the whole database to find the principle feature components and then map each audio clip into one vector in the principle feature space.

3 Dominant Feature Vector based Similarity Measure

Based on the extracted dominant feature vectors, a new similarity measure is correspondingly proposed in this section. The characteristics of the measure are discussed and the comparisons with other conventional methods are also given.

3.1 Similarity Measure Definition

Consider two audio clips which contain m_1 and m_2 dominant feature vectors respectively, their i^{th} and j^{th} dominant feature vectors is denoted as q_i and p_j , and the corresponding eigen-value is λ_i and σ_j . To measure the similarity between these two audio clips, the similarity between q_i and p_j is firstly considered, which is usually defined as their inner-product:

$$s_{i,j} = \frac{\|q_i^T p_j\|^2}{\|q_i\|^2 \|p_j\|^2} = \|q_i^T p_j\|^2 \quad (5)$$

Since different dominant feature vector has different importance, which is determined by the corresponding eigen-values, in representing an audio clip, they should have different contributions to the audio similarity measure. That is to say, the dominant feature vectors with large eigen-values should contribute more in similarity measuring between two audio clips. Thus, the similarity of two audio clips is defined as the weighted sum of the similarity between every two of their dominant feature vectors:

$$S = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} w_{i,j} s_{i,j} \quad (6)$$

where the weighting factor $w_{i,j}$ is determined by the corresponding eigen-values:

$$w_{i,j} = \frac{\lambda_i}{\sqrt{\sum_{i=1}^{m_1} \lambda_i^2}} \frac{\sigma_j}{\sqrt{\sum_{j=1}^{m_2} \sigma_j^2}} \quad (7)$$

The weighting factor is such chosen for the following two considerations: 1) it should be proportional to the contributions of the corresponding dominant

feature vectors q_i and p_j ; and 2) This weighted sum should be equal to one, when two audio clips are the same, i.e., $q_i = p_j$ and $\lambda_i = \sigma_j$.

Actually, the dominant feature vectors are obtained from the covariance matrix of the frame based feature vector, and construct the base of the signal subspace. From this point of view, it can be considered that the similarity between two clips is in essence measured based on their noise-reduced covariance matrices.

3.2 Properties of the Similarity Measure

As mentioned above, the similarity between two clips is measured based on their signal subspaces, which can be obtained from the covariance matrices. Therefore, their similarity is actually a function of C_1 and C_2 , denoted as $S(C_1, C_2)$, where C_1 and C_2 are two covariance matrices of two clips, respectively. The properties of this similarity measure are discussed in this section.

Firstly, the similarity is symmetric when comparing two covariance matrices:

$$S(C_1, C_2) = S(C_2, C_1) \quad (8)$$

It is a basic requirement for most of similarity measure or distance measure.

Secondly, the similarity measure is normalized to the range from 0 to 1:

$$0 \leq S(C_1, C_2) \leq 1 \quad (9)$$

A larger value indicates more similar between the two clips. If two clips have the same dominant feature vectors and the same corresponding eigen-values, their similarity will be 1. Otherwise, if their dominant feature vectors are totally different, that is, orthogonal with each other, the similarity will be 0.

Moreover, the proposed similarity measure is robust. For example, if an audio clip is contaminated by superimposed or sequentially concatenated noise, its statistical features including mean and standard deviation will be affected much. However, its salient characteristics, or the corresponding dominant feature vectors, will not have significant difference. Therefore, our proposed similarity measure is not sensitive to the effect of noise.

For a better understanding of the proposed similarity measure, we explain it in a more intuitive way. As well known, one covariance matrix C determines a corresponding hyper-ellipse:

$$\{x : x^T C^{-1} x = 1\} \quad (10)$$

Assuming that q_i is one of the eigen-vector and λ_i is the corresponding eigen-value, thus q_i determines the orientation of one semi-axis of the hyper-ellipse, $\sqrt{\lambda_i}$ measures the length of the corresponding axis. The shape and orientation of the hyper-ellipse is an intuitive description of the characteristics of an audio clip. Fig. 2 illustrates two different hyper-ellipses in the 2-dimensional plane as an example.

In this way, the similarity between two audio clips can also be treated as the similarity of the shape and orientation between two corresponding hyper-ellipses. Thus, our proposed similarity measure can be deduced consequently, since the similarity between the orientations is measured by the inner-product of different dominant feature vectors, and the similarity between the shapes is affected by the weighting factor defined in Eq. (7).

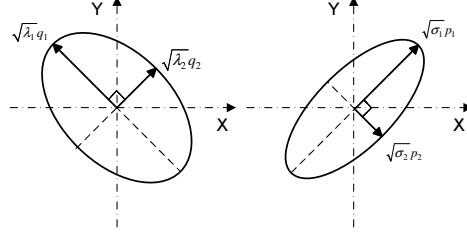


Fig. 2. Illustrations of two 2-D hyper-ellipses with two dominant feature vectors, a “fat” one is on the left, and a “slender” one is on the right.

3.3 Comparison with Other Distance Measures

In previous researches, several distance measures have been proposed and studied based on two covariance matrix [4], such as Kullback-Leibler in Eq. (11) and the Bhattacharyya in Eq. (12).

$$d_{KL}(C_1, C_2) = \frac{1}{2}(\bar{\mu}_2 - \bar{\mu}_1)^T (C_2^{-1} - C_1^{-1})(\bar{\mu}_2 - \bar{\mu}_1) + \frac{1}{2} \text{tr}(C_1^{-1}C_2 + C_2^{-1}C_1 - 2I) \quad (11)$$

$$d_{BHA}(C_1, C_2) = \frac{1}{4}(\bar{\mu}_2 - \bar{\mu}_1)^T (C_2^{-1} - C_1^{-1})(\bar{\mu}_2 - \bar{\mu}_1) + \frac{1}{2} \log \frac{\|C_1 + C_2\|}{2\sqrt{\|C_2C_1\|}} \quad (12)$$

where $\bar{\mu}$ is the mean of the sample vectors, and C_1 and C_2 are two covariance matrixes.

However, these distances utilize the inverse of covariance, and are usually used for the similarity measure between two sets of data, where a covariance matrix can be accurately estimated from sufficient data and represents the feature distribution of corresponding data set. In general, these distances are not suitable in measuring the distance between two audio clips, because of the following problems:

1. The covariance estimated from an audio clip is easily affected by noise.
2. If the duration of an audio clip is sometimes short and does not have enough sample frames, the estimated covariance matrix may not be full rank or ill conditioned. This will leads to numerical instability.
3. The noise-reduced covariance matrix can not be used directly in Eq. (11) and Eq. (12), since it is usually not full rank and thus not invertible.

The proposed similarity measure does not have these problems. Even if the original covariance matrix is not full rank, we can still extract the dominant feature vectors.

4 Experiments

In order to demonstrate the effectiveness of the proposed similarity measure, we compared its performance with some other similarity or distance measure, including L_2 distance, Kullback-Leibler distance and Bhattacharyya distance, based on a content-based audio retrieval system.

Our testing database consists of around 1000 audio clips. These sounds vary in duration from less than one second to about 30 seconds; and include many kinds of sounds, such as *animals, machines, vehicles, human, weapons* and so on.

In our experiment, all audio streams are down-sampled into 8 KHz, 16-bit and mono-channel, for universal processing. Each frame is of 200 samples (25ms), with 50% overlapping. Two types of features are computed for each frame: (i) perceptual features and (ii) 8 order Mel-frequency Cepstral Coefficients (MFCC). The perceptual features are composed of short time energy, zero crossing rate, pitch, 8 order subband energies, brightness and bandwidth. These features are then combined as a 21-dimensional feature vector for a frame.

The first experiment is to find the best threshold η defined in Eq. (3), in order to decide how many dominant feature vectors should be used in the similarity measure. Fig. 3 illustrates a representative curve of recall ratio at the top 20 selection, indicating the influence of the threshold η from 80% to 100%. It can be seen that with the threshold increases, performance improves at first. The performance almost stops improving or even decreases when the threshold is more than 90%. The reason is that the first several dominant feature vectors contain important information of the clip, while the remaining is formed by noise effect and degrades the performance. In the following experiment, η is set as 90% to select the number of dominant feature vectors.

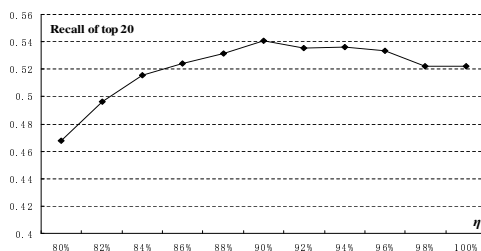


Fig. 3. The performance of the proposed similarity measure when the number of dominant feature vectors increase.

Fig. 4 illustrates the comparison results among the proposed similarity measure, L_2 , Kullback-Leibler and Bhattacharyya distance. In the experiments, L_2

distance is based on the mean and standard deviation of frame features of a clip; Kullback-Leibler and Bhattacharyya distance are both based on the covariance matrix. From Fig. 4, it can be seen that the proposed method has an obvious

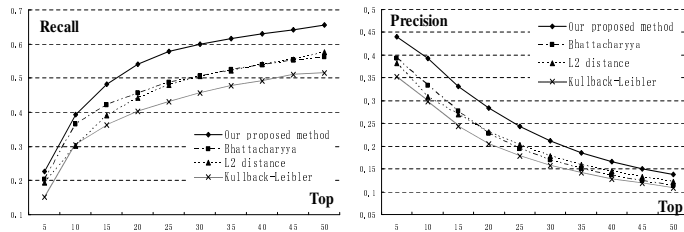


Fig. 4. Comparisons among the proposed similarity measure, L_2 distance, Kullback-Leibler and Bhattacharyya distance.

improvement, compared with the other similarity measures. For example, in the results of top 20, about 55% targets are retrieved with the proposed method, while only 45% is obtained using common L_2 distance. The corresponding results using Kullback-Leibler and Bhattacharyya are about 46% and 40% respectively. The precision is also increased compared with conventional similarity measures. In the results of top 20, the precision of the proposed method is about 28%, while other measure methods are less than 23%. The improvement is about 22%.

5 Conclusion

In this paper, a new similarity measure between audio clips is proposed. This similarity measure is based on the dominant feature vectors extracted from an audio clip. Compared with conventional mean and standard deviation, the dominant feature vectors give a better representation of an audio clip, especially when the characteristics change with time. Experimental results demonstrate the effectiveness of the proposed similarity measure. It also indicates that our approach is better than conventional L_2 distance, Kullback-Leibler distance and Bhattacharyya distance, in the case of similarity measure of two audio clips.

References

1. L. Lu, H.-J. Zhang, and H. Jiang, "Content Analysis for Audio Classification and Segmentation", *IEEE Trans. on Speech and Audio Processing*, 10(7):504-516, 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", *Proc. of 11th ACM Multimedia*, pp. 219-222, 2003.
4. M. Basseville, "Distance Measure for Signal Processing and Pattern Recognition", *Signal Processing*, 18(4):349-369, 1989.