International Journal of Pattern Recognition
and Artificial Intelligence
Vol. 19, No. 6 (2005) 823–837
© World Scientific Publishing Company

CONTENT-BASED AUDIO RETRIEVAL BASED ON GABOR WAVELET FILTERING*

RUEI-SHIANG LIN and LING-HWEI CHEN[†]

Department of Computer and Information Science
National Chiao Tung University, 1001 Ta Hsueh Rd.
Hsinchu, Taiwan 30050, R.O.C.
[†]*lhchen@cc.nctu.edu.tw*

Rapid increase in the amount of audio data and especially music collections demand an efficient method to automatically retrieve audio objects based on its content. In this paper, based on the Gabor wavelet features, we will propose a method for content-based retrieval of perceptually similar music pieces in audio documents. It allows the user to select a reference passage within an audio file and retrieve perceptually similar passages such as repeating phrases within a music piece, similar music clips in a database or one song sung by different persons or in different languages.

The proposed method will first divide an audio stream into clips, each of which contains one-second audio information. Then, the frame-based features of each clip are extracted based on the Gabor wavelet filters. Finally, a similarity measuring technique is provided to perform pattern matching on the resulting sequences of feature vectors. Experimental results show that the proposed method can achieve over 96% accuracy rate for audio retrieval.

Keywords: Spectrogram; audio content-based retrieval; Gabor wavelets; singular value decomposition.

1. Introduction

The recent emergence of multimedia and the tremendous growth of multimedia data archives have made the effective management of multimedia databases a very important and challenging task. Therefore, developing an efficient searching and indexing technique for multimedia databases becomes very important and have drawn lots of attention recently. As considerable research work has been done on the content-based retrieval of image and video data, less attention has been received for the content-based retrieval of audio data.

*This research was supported in part by the Department of Industrial Technology of R.O.C. under contract 93-EC-17-A-02-S1-032 and the National Science Council of R.O.C. under contract NSC-92-2213-E-009-101.

[†]Author for correspondence.

In recent years, techniques for audio information retrieval have started emerging as research prototypes.[2,3,5,6,8,9,13,15,18,20–22,24–29] These systems can be classified into two major paradigms.[11,21] In the first paradigm, the user sings a melody and similar audio files containing that melody are retrieved. This kind of approach[5] is called "Query by Humming" (QBH). It has the disadvantage of being applicable only when the audio data is stored in symbolic form such as MIDI files. The conversion of generic audio signals to symbolic form, called polyphonic transcription, is still an open research problem in its infancy.[21] Another problem with QBH is that it is not applicable to several musical genres such as dance music where there is no singable melody that can be used as a query. The second paradigm[2,18,20,21,24,26,27] is called "Query-by-Example" (QBE), a reference audio file is used as the query and audio files with similar content are returned and ranked by their similarity. In order to search and retrieve general audio signals such as the raw audio files (e.g. mp3, wave, etc.) on the web or databases, only the QBE paradigm is currently applicable. In this paper, we will develop a QBE system that will work directly on real world raw audio data without attempting to transcribe the music.

Wold *et al.*[24] proposed an approach to retrieve the audio objects based on their content in waveform. In this approach, an N-vector for a given sound is constructed according to the acoustical features including loudness, pitch, brightness, bandwidth and harmonicity. The N-vector is then used to classify sounds for similar searching. This method is only suitable for sounds with a single timbre. Besides, the method is supervised and not adequate to index general audio content. An approach based on the histogram model of the zero-crossing features for searching quickly through broadcast audio data was provided in Ref. 18. In this approach, a certain reference template is defined and applied on each audio stream to find whether it contains the desired reference sound. The accuracy of the result using this method varies considerably for different types of recording. Besides, the audio segment to be searched should be known *a priori* in this algorithm.

Foote[2] proposed a data driven approach for audio data retrieval by computing the Mel-frequency cepstral coefficients (MFCCs) of an audio signal first. Then a learning algorithm is applied on these MFCCs to generate a quantization tree. Each kind of audio signals is inserted into the corresponding bin in the quantization tree. Cosine measurement or Euclidean distance can be used to measure the similarity between two bins. A QBE system called "SoundSpotter"[20] provides a sound classification tool to classify a large database into several categories and finds the best matches to the selected query sound using state-path histograms. It is also based on the MFCCs representation. Both of the above-mentioned two MFCC-based approaches are not suitable for melody retrieval (e.g. music) since the MFCC-based features do not capture enough information about the pitch content, rather, they characterize the broad shape of the spectrum. In Ref. 27, local peaks in spectrogram are identified and a spectral vector is extracted near each peak. Since the parameters used in the peak identification algorithm are too many and empirical, they are improper when the source of audio signals is changed.

In this paper, based on the Gabor wavelet features, we will propose a method for content-based retrieval of perceptually similar music pieces in audio documents. It is based on the QBE paradigm and allows the user to select a reference passage within an audio file and retrieve perceptually similar passages such as repeating phrases within a music piece, similar music clips in a database or one song sung by different persons or in different languages. The proposed method consists of four phases: time-frequency distribution (TFD) generation, initial feature extraction, feature selection and similarity measurement. First, the input audio stream is transformed to a spectrogram and divided into clips, each of which contains one-second audio information and will meet the human auditory system (HAS).[30] Second, for each clip with one-second window, a set of initial frame-based features are extracted based on the Gabor wavelet filters.[4,10] Third, based on the extracted initial features, the Singular Value Decomposition (SVD)[1] is used to perform the feature selection and to reduce the feature dimension. Finally, a similarity measuring technique is provided to perform pattern matching on the resulting sequences of feature vectors.

Experimental results show that the proposed method can achieve over 96% accuracy rate for audio retrieval and the complexity is low enough to allow operation on today's personal computers and other cost-effective computing platforms. These results demonstrate the capability of the proposed audio features for characterizing the perceptual content of an audio sequence. The paper is organized as follows. In Sec. 2, the proposed method will be described. Experimental results will be presented in Sec. 3. Finally, the conclusions will be given in Sec. 4.

2. The Proposed System

The block diagram of the proposed method is shown in Fig. 1. It is based on the spectrogram and consists of four phases: time-frequency distribution (TFD) generation, initial feature extraction, feature selection and similarity measurement. First, the input audio is transformed to a spectrogram, which will meet the human auditory system (HAS).[30] Second, for each clip with one-second window, some Gabor wavelet filters will be applied to the resulting spectrogram to extract a set of initial features. Third, based on the extracted initial features, the Singular Value Decomposition (SVD)[1] is used to perform the feature selection and to reduce the feature dimension. Finally, based on the selected features, a similarity measure is provided to measure the similarity of audio data. In what follows, we will describe the details of the proposed method.

2.1. *TFD generation*

In the first phase, the input audio is first transformed to a spectrogram that is a commonly used representation of an acoustic signal in a three-dimensional (time, frequency, intensity) space known as a time-frequency distribution (TFD).[16] Conventionally, the Short Time Fourier Transform (STFT) is applied to construct a spectrogram and the TFD is sampled uniformly in time and frequency. However,

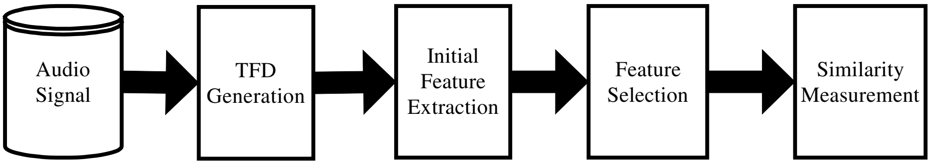


Fig. 1. Block diagram of the proposed method.

this is not suitable for the auditory model because the frequency resolution within the human psycho-acoustic system is not constant but varies with frequency.[30]

In this paper, the TFD is perceptually tuned, mimicking the time-frequency resolution of the ear. That is, the TFD consists of axes that are nonuniformly sampled. Frequency resolution is coarse and temporal resolution is fine at high frequencies while temporal resolution is coarse and frequency resolution is fine at low frequencies.[30] Given the sampling frequency (Fs) of 44100 Hz, the Hamming window is applied and an audio signal is divided into frames, each of which contains 512 samples ($N = 512$), with 50% overlap in every two adjacent frames. One example of the tiling in the time-frequency plane is shown in Fig. 2. Figure 3 shows a schematic diagram of the TFD generation.

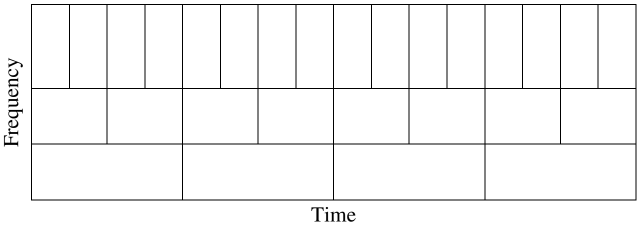


Fig. 2. An example of tiling in the time-frequency plane.

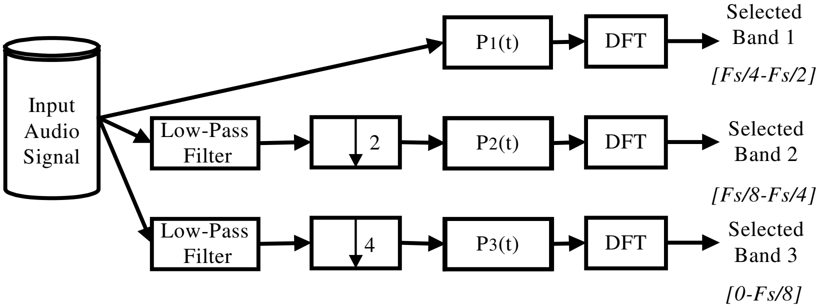


Fig. 3. A schematic diagram of the TFD generating details.

There are three parts in the TFD generation. In the first part, the N-point DFT (Discrete Fourier Transform) is applied to the original audio signal $P_1(t)$ to obtain a spectrogram $S_1(x, y)$. In the second part, a low-pass filter is first applied to $P_1(t)$ and then the filtered result is downsampled half-size to obtain signal $P_2(t)$ and the N-point DFT is applied to $P_2(t)$ to obtain a spectrogram $S_2(x, y)$. In the third part, a low-pass filter is first applied to $P_1(t)$ and then the filtered result is downsampled quarter size to obtain a signal $P_3(t)$ and the N-point DFT is applied to $P_3(t)$ to obtain a spectrogram $S_3(x, y)$. Note that the downsampling is conducted after applying a low-pass filtering to the original signal to prevent the aliasing, and the window size for DFT is 512 (i.e. $N = 512$) in this paper. The frequency resolution Δf_j and the analysis time interval T_j in $S_j(x, y)$ can be calculated as follows:

$$\Delta f_j = \frac{1}{2^{j-1}} \cdot \frac{FS}{N} = \frac{1}{T_j}, \quad j = 1,\, 2\,, 3. \tag{1}$$

Note that the window center at the kth time block in $S_j(x, y)$, t_j^k, is given by

$$t_j^k = \frac{k}{2} T_j, \quad j = 1, 2, 3. \tag{2}$$

Finally, based on $S_1(x, y)$, $S_2(x, y)$, and $S_3(x, y)$, a spectrogram $I(x, y)$ is obtained according to the following equation:

$$I(x,y) = \begin{cases} S_1(x,y), & \text{if } y \in [F_S/4, F_S/2], \\ & \quad x = 0, 1, \ldots, N_f - 1; \\ S_2(2i,y), & \text{if } y \in [F_S/8, F_S/4], \\ & \quad x = 2i, 2i+1, \quad i = 0, 1, \ldots N_f/2 - 1; \\ S_3(4i,y), & \text{if } y \in [0, F_S/8], \\ & \quad x = 4i, \ldots, 4i+3, \quad i = 0, 1, \ldots N_f/4 - 1; \end{cases} \tag{3}$$

where N_f is the frame number of $P_1(t)$. From Eq. (3), we can see that in $I(x, y)$, frequency resolution is coarse and temporal resolution is fine at high frequencies while temporal resolution is coarse and frequency resolution is fine at low frequencies. This means that $I(x, y)$ meets the human psycho-acoustic system.

2.2. *Initial feature extraction*

Generally speaking, the spectrogram is a good representation for the audio since it is often visually interpretable. By observing a spectrogram, we can find that the energy is not uniformly distributed, but tends to cluster to some patterns.[17] All curve-like patterns are called tracks. Figure 4(a) shows that for a musical instrument signal, some line tracks corresponding to tones will exist on its spectrogram. Figure 4(b) shows some patterns including clicks (broadband, short time), noise burst (energy spread over both time and frequency), tones, and frequency sweeps in a song spectrogram. Thus, if we can extract some features from a spectrogram to represent these patterns, the retrieval should be easy. Smith and Serra[19] proposed a method to extract tracks from a STFT spectrogram. Once the tracks are extracted, each track is classified. However, tracks are not well suited for describing some kinds of patterns such as clicks, noise burst and so on. To treat all kinds of patterns, a richer representation is required. In fact, these patterns contain various orientations and spatial scales. For example, each pattern formed by lines [see Fig. 4(a)] will have a particular line direction (corresponding to orientation) and width (corresponding to spatial scale) between two adjacent lines; each pattern formed by curves [see Fig. 4(b)] contains multiple line directions and a particular width between two neighboring curves. Since Gabor wavelet transform provides an optimal way to extract those orientations and scales,[16] in this paper, we will use the Gabor wavelet functions to extract some initial features to represent the needed patterns. Note that in this paper, we will only deal with musical audio signal (musical instrument, song, etc.). The detail will be described in the following section.

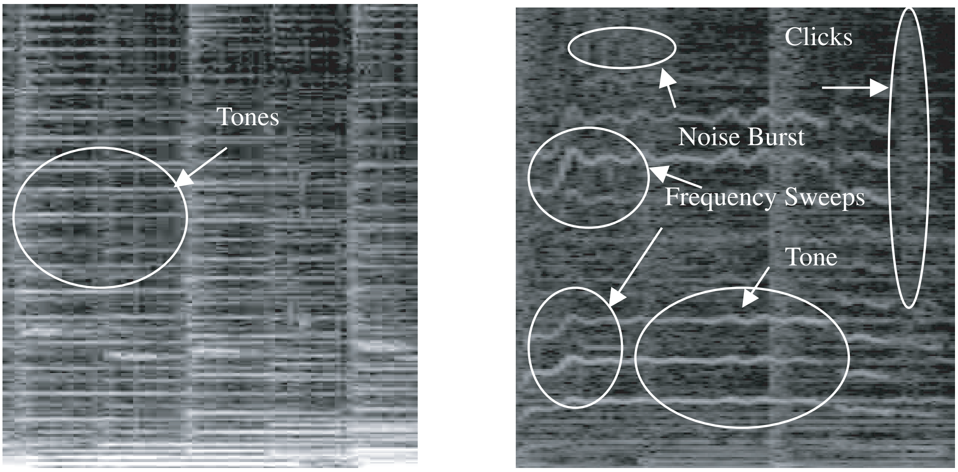
2.2.1. *Gabor wavelet functions and filters design*

Two-dimensional Gabor kernels are sinusoidally modulated Gaussian functions. Let $g(x, y)$ be the Gabor kernel, its Fourier Transform $G(u, v)$ can be defined as follows[10]:

$$g(x,y) = \left(\frac{1}{2\pi\sigma_x\sigma_y}\right)\exp\left[\frac{-1}{2}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}\right)+2\pi j\omega x\right], \tag{4}$$

$$G(u,v) = \exp\left(\frac{-1}{2}\left[\frac{(u-\omega)^2}{\sigma_u^2}+\frac{v^2}{\sigma_v^2}\right]\right), \tag{5}$$

where $\sigma_u = \frac{1}{2\pi\sigma_x}$ and $\sigma_v = \frac{1}{2\pi\sigma_y}$ and ω is the center frequency.



(a)

(b)

Fig. 4. Two examples to show some different possible kinds of patterns in a spectrogram. (a) Line tracks corresponding to tones in a musical instrument spectrogram. (b) Clicks, noise burst, tones and frequency sweeps in a song spectrogram.

Gabor wavelets are sets of Gabor kernels which will be applied to different subbands with different orientations. It can be obtained by appropriate dilations and rotations of $g(x, y)$ through the following generating functions^{10}:

$$g_{mn}(x, y) = a^{-m} g(x', y'), \quad a > 1, m, n = \text{integer},$$

$$x' = a^{-m}(x \cos\theta + y \sin\theta), \quad \text{and} \quad y' = a^{-m}(-x \sin\theta + y \cos\theta), \tag{6}$$

$$a = \left(\frac{\omega_h}{\omega_l}\right)^{\frac{1}{S-1}}, \tag{7}$$

$$\sigma_u = ((a-1)\,\omega_h)/((a+1)\sqrt{2\ln 2}), \tag{8}$$

$$\sigma_v = \tan\left(\frac{\pi}{2k}\right)\left[\omega_h - 2\ln 2\left(\frac{\sigma_u^2}{\omega_h}\right)\right]\left[2\ln 2 - \frac{(2\ln 2)^2 - \sigma_u^2}{\omega_h^2}\right]^{-\frac{1}{2}}, \tag{9}$$

where $\theta = \frac{n\pi}{K}$, $n = 0, 1, \ldots, K-1$, $m = 0, 1, \ldots, S-1$, K is the total number of orientations, S is the number of scales in the multiresolution decomposition, ω_h and ω_l are the lowest and the highest center frequencies, respectively.

In this paper, we will deal with musical audio signal including musical instrument and song. Most of the current works only deal with the monophonic sources, in this paper we will also consider polyphonic music. Polyphonic music is more common, but it is also more difficult to represent. The most meaningful feeling of human perception for the music data is primarily the pitch and timbre. Both of them are correlated with the tones. For example, the fundamental tone decides the pitch that we hear, and the harmonics decide the timbre. Based on the above-observation for the spectrogram [see Figs. 4(a) and 4(b)], we find that some line tracks corresponding to tones will exist in the spectrogram. Thus, if we can extract the features about tones, the retrieval should be easy. Note that in our experiments, we set $\omega_l = 3/64$, $\omega_h = 3/4$, $K = 1$ and $S = 7$.

2.2.2. *Feature estimation*

Since through our observation, most prominent tracks are near horizontal, in this paper, we only take one orientation that is horizontal. Thus, each Gabor wavelet filter, $g_{mn}(x, y)$, can be briefly represented by $g_m(x, y)$. To extract the audio features, each Gabor wavelet filter, $g_m(x, y)$, is first applied to the spectrogram $I(x, y)$ to get a filtered spectrogram, the spectrum of which is represented by $W_m(u, v)$ called spectrogram spectrum. In this paper, the above filtering process is executed in frequency domain through the following equation:

$$W_m(u, v) = F\{g_m(x, y)\} \cdot F\{I(x, y)\}, \tag{10}$$

where $F\{\cdot\}$ is a fast Fourier Transform. Up to now, there are S spectrogram spectrum with scale m, $W_m(u, v)$, to be available. Since, in each audio signal, those tracks appearing in the corresponding spectrum have a certain scale, not all these spectrogram spectrum are used to perform the feature estimation, only the one with the maximum contrast (which corresponds to the track scale) is used. To reach this goal, the vertical profile of the spectrum, $P_m(u)$ $(m = 1, 2, \ldots, S)$, is constructed as follows:

$$P_m(u) = \sum_v W_m(u, v). \tag{11}$$

Let M_P be the number of local peaks $(u_1, u_2, \ldots, u_{M_P}.)$ in $P_m(u)$, $P_m(u_i)$ $(i = 1, 2, \ldots, M_P)$ be the magnitudes of these peak points, and

$$P_m^{\max} = \max_{u_i} P_m(u_i). \tag{12}$$

Then the contrast is defined as

$$\text{contrast}_m = P_m^{\max} - \frac{1}{M_P} \sum_{i=1}^{M_P} P_m(u_i). \tag{13}$$

Let

$$mc = \underset{m}{\arg}\ \text{contrast}_m, \tag{14}$$

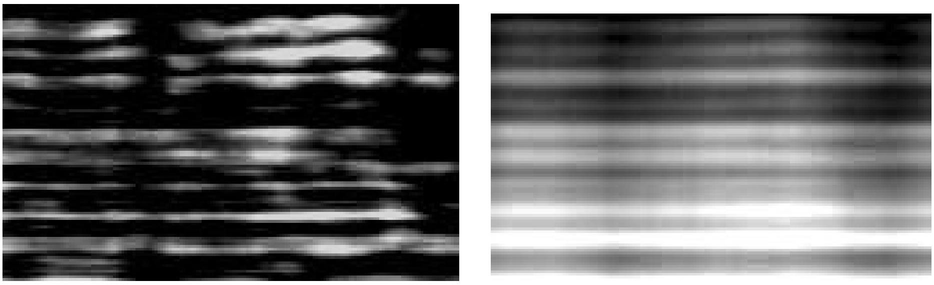
then the spectrogram spectrum, $W_{mc}(u, v)$, and the corresponding spectrogram, $w_{mc}(x, y)$, are used for initial feature extraction.

Figure 5(a) shows an example of the Gabor-wavelet filtered spectrogram with the maximum contrast, $w_{mc}(x, y)$. From Fig. 5(a), we can see that the tracks in the figure are somewhat obscured, to remove this phenomenon, an enhancement process[23] is applied as follows:

$$w_f(x, y) = F^{-1} \left\{ W_{mc}(u, v) \cdot |W_{mc}(u, v)|^{\alpha} \right\}, \tag{15}$$

where α is set as 1.4 and $w_f(x, y)$ is the enhanced spectrogram. Figure 5(b) shows the result of the enhancement process in Fig. 5(a).

An initial feature vector, $\mathbf{f}$, is now constructed using $w_f(x, y)$ as feature components. Recall that in our experiments, for each clip, one-second window (M frames) is used for constructing spectrogram. Besides, high frequency components above



(a)

(b)

Fig. 5. An example to show the enhancement process performed in a spectrogram. (a) The Gabor-wavelet filtered spectrogram with the maximum contrast. (b) Enhanced spectrogram.

$Fs/4$ are discarded to avoid the influence of noise. These will result in a $M \times N$-dimensional initial feature vector

$$\mathbf{f} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M]^t, \tag{16}$$

where $\mathbf{x}_i = [w_f(i,1), w_f(i,2), \ldots, w_f(i,N)]$ $(i = 1, 2, \ldots, M)$ is the spectral vector of each frame in $w_f(x,y)$.

2.3. *Feature selection and representation*

The initial features are not used directly for similarity measurement since some features give poor separability among different objects and inclusion of these features will lower down the system performance. In addition, some features are highly correlated so that redundancy will be introduced. To remove these disadvantages, in this paper, the Singular Value Decomposition (SVD)[1] is applied to the initial features to find those uncorrected features with the highest separability.

As for the SVD, it is a well-known technique for reducing the dimensionality of data while retaining maximum information content. It decomposes the data into a sum of vector outer products with vectors representing both the basis function (eigenvectors) and the projected features (eigen coefficients). A subset of the complete basis is selected to reduce data dimensionality. The loss of information is minimized because the basis functions are ordered by statistical salience; thus, functions with low information content are discarded.

Based on SVD, the initial feature vector, $\mathbf{f}$, for each one-second audio clip can be decomposed into the form[7]:

$$\mathbf{f} = \mathbf{USV}^t, \tag{17}$$

where $\mathbf{S}$ is a diagonal matrix containing the singular values of $\mathbf{f}$ along its diagonal, and the columns of $\mathbf{U}$ and $\mathbf{V}$ are the eigenvectors (the basis function) of $\mathbf{f}\,\mathbf{f}^t$, and $\mathbf{f}^t\mathbf{f}$ respectively. Then the basis, $\mathbf{V}$, is reduced by retaining only the first k basis functions. That is

$$\mathbf{V}_k = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k]. \tag{18}$$

And the initial feature vector $\mathbf{f}$ is projected to the space generated by $\mathbf{V}_k$ to get a new feature vector $\mathbf{f}'$ with the reduced dimension. $\mathbf{f}'$ is then used to stand for the audio clip as follows:

$$\mathbf{f}' = [\mathbf{x}'_1, \mathbf{x}'_2, \ldots, \mathbf{x}'_M]^t = \mathbf{f}\,\mathbf{V}_k, \tag{19}$$

where $\mathbf{x}'_i$ $(i = 1, 2, \ldots, M)$ is a k-dimensional vector. Note that we will call $\mathbf{V}_k$ as the basis of $\mathbf{f}'$.

2.4. *Audio retrieval and similarity measurement*

In general, audio (multimedia) data searching can be classified into two different strategies: "a-whole-object search", and "in-object search". "A-whole-object search" approach searches for data that is globally similar to the query input; on the other hand, an "in-object search" approach searches for a large piece of data containing a fragment that is similar to the query. A method of using the latter searching strategy can reach the aim of the first searching strategy but not vice versa. Thus, in this paper, the retrieval is performed based on the latter searching strategy. Based on the feature vector introduced in the previous section, the similar audio clip retrieval will be conducted. Before retrieval, it is important to give a good similarity measure. Here, a distance measure is first proposed to evaluate the similarity between two audio clips. In our experiments, the Euclidean distance worked better than others (e.g. Mahalanobis, covariance, etc.) in the space generated by $\mathbf{V}_k$.

2.4.1. *Similarity measure*

For the candidate audio sequence, $\mathbf{y}_c$ with feature vector $\mathbf{f}'_j = [\mathbf{x}'_{j,1}, \mathbf{x}'_{j,2}, \ldots, \mathbf{x}'_{j,M}]$ $(j = 1, 2, \ldots, l)$, where l is the number of one-second clips in the audio sequence. That is, $\mathbf{y}_c$ is divided into one-second clips:

$$\mathbf{y}_c = [y_1, y_2, \ldots, y_l], \tag{20}$$

where y_j has feature vector $\mathbf{f}'_j$.

For every queried one-second clip, y_q, before computing the distance between y_q and each of the candidate clip y_j, y_q should be projected to the basis of y_j to get the corresponding feature $\mathbf{f}'_q = [\mathbf{x}'_{q,1}, \mathbf{x}'_{q,2}, \ldots, \mathbf{x}'_{q,M}]^t$. Then the distance between one-second clips y_q and y_j is evaluated as follows:

$$\mathrm{Dist}_{q,j} = \left(\sum_{i=1}^{M} \left| \mathbf{x}'_{q,i} - \mathbf{x}'_{j,i} \right|^2 \right)^{\frac{1}{2}}, \tag{21}$$

where $j = 1, 2, \ldots, l$ and $\left| \mathbf{x}'_{q,i} - \mathbf{x}'_{j,i} \right|$ stands for the Euclidean distance between two vectors: $\mathbf{x}'_{q,i}$ and $\mathbf{x}'_{j,i}$. Then for all j, sort $\mathrm{Dist}_{q,j}$ in an increasing order. For the top g clips, we define their grades, $Gd_{q,j}$, as $g, g-1, g-2, \ldots,$ and 1, respectively. The clip with the least distance will have the highest grade and be considered as the most similar one. In addition, $Gd_{q,j}$ of all other clips are defined as zero. Note that in this paper, one-second audio clip is taken as the basic distance measurement unit.

2.4.2. *Retrieval*

For a query audio sequence, $\mathbf{y}_q$, with length p-seconds, it is first divided into p successive one-second clips. That is

$$\mathbf{y}_q = [y_q^1, y_q^2, \ldots, y_q^p]. \tag{22}$$

Next, for each clip y_q^i $(i = 1, 2, \ldots, p)$ and a candidate audio sequence $\mathbf{y}_c$, the similarity measure is first performed and the corresponding grades, $Gd_{q,j}^i$

($i = 1, 2, \ldots, p$ and $j = 1, 2, \ldots, l$), are evaluated based on Eq. (21). According to these grades, the total grade of the clip y_q^i, Gd_T_{q,d_i}^i, is defined to be

$$\begin{aligned} Gd\_T_{q,d_i}^i &= Gd_{q,d_i-i+1}^1 + \cdots + Gd_{q,d_i-1}^{i-1} + Gd_{q,d_i}^i + Gd_{q,d_i+1}^{i+1} \\ &\quad + \cdots + Gd_{q,d_i+p-i}^p, \end{aligned} \tag{23}$$

where $d_i = \arg_j \max Gd_{q,j}^i$ and is the candidate for matching.

Finally, based on the set of total grades, Gd_T_{q,d_i}^i ($i = 1, 2, \ldots, p$), the global similarity is defined as

$$\mathrm{Sim} = 1 - \frac{\max DGd\_T_{q,c}^i}{\sum_{i=1}^{p-1} DGd\_T_{q,c}^i}, \tag{24}$$

where $DGd\_T_{q,d_i}^i = Gd\_T_{q,d_i}^i - Gd\_T_{q,d_i}^{i+1}$ ($i = 1, 2, \ldots, p-1$). If the global similarity, Sim, is less than a predefined threshold, mark the query sequence as ambiguous and no query result will be available. Otherwise, the matching clip with the highest similarity to the query can be retrieved according to the following criterion:

$$(s, r) = \underset{(i,d_i)}{\arg} \max Gd\_T_{q,d_i}^i, \tag{25}$$

where $i = 1, 2, \ldots, p$, and the best matched audio sequence, $\mathbf{y}_o$, in the candidate audio sequence will result in the following audio sequence:

$$\mathbf{y}_o = [y_{r-s+1}, \ldots, y_{r-1}, y_r, y_{r+1}, \ldots, y_{r+p-s}]. \tag{26}$$

Besides, based on Eq. (25), if the top n matched audio sequence with total grades, Gd_T_{q,d_i}^i, are larger than a predefined threshold, they can be considered as the repeating audio sequence of the query.

3. Experimental Results

3.1. *Audio database*

In order to show the efficiency of the proposed method, we have collected a set of 150 musical pieces (50 musical instruments, 100 songs) with total length about ten hours and more than 10,000 phrases as the testing database. Care was taken to obtain a wide variation in each type such as varied instruments, different languages (English, Chinese, Japanese, etc.), different singers (male, female, or children), and different style (jazz, rock, folk, etc.). These audio clips are stored as 16-bit per sample with 44.1 kHz sampling rate in the WAV file format and are used to test the audio retrieval performance. Note that in order to compare, the testing database includes the dataset described in Refs. 26 and 27, and some of the clips are taken from MPEG-7 content set.[14]

3.2. *Experiment results*

There are two major factors affecting the performance of the proposed approach, i.e. the number of basis functions used and the length of the query example. In order to examine the performance of the proposed method, we present two experiments. In the first experiment, for each music object in the database, we use its refrain as the query example to retrieve all repeating phrases similar to this refrain. Therefore, 150 queries are performed. This experiment is presented to examine the quality of the proposed retrieval approach with two above-mentioned major factors. As for the second experiment, for each song, there will be two versions which are sung in different languages or by different persons in the database. We use its refrain in a certain version (e.g. the Chinese version) as the query example to retrieve all repeating phrases similar to this refrain in another version (e.g. the English version). This experiment is presented to examine the robustness of the proposed retrieval approach.

In this paper, the performance is evaluated by the precision rates (P_r) and the recall rates (R_e).[12] Note that the recall rate, R_e, and the precision rate, P_r, are defined as

$$R_e = \frac{N}{T} \quad \text{and} \quad P_r = \frac{N}{K}, \tag{27}$$

where N is the number of relevant items retrieved (i.e. correctly retrieved items), T is the total number of relevant items (i.e. correctly retrieved items and the relevant items that have not been retrieved) and K is the total number of retrieved items. The recall rate is typically used in conjunction with the precision rate, which measures the fraction of the retrieved patterns that is relevant. The precision and recall rate can often be traded-off. That is one can achieve high precision rate and low recall rate or the other way round.

Tables 1 and 2 show the results of two experiments presented in this paper. In our experiments, the number of retrieved patterns was adjusted to the number of relevant patterns, so the precision and recall rates are the same.

From Table 1, we can see that the above-mentioned two factors affect the performance of the proposed approach. The more basis functions are used, the higher the recall rate will be. And the longer the length of the query sample used, the higher the recall rate will be. Based on the first experiment, we can see that it is

Table 1. The average recall rates of the first experiment.

Basic Function Numbers	Query Sample Length		
	One Second	Two Seconds	Three Seconds
5	29%	71%	74%
10	31%	75%	75%
15	40%	98%	98%

Table 2. The average recall rates of the second experiment.

Basic Function Numbers	Query Sample Length		
	One Second	Two Seconds	Three Seconds
5	31%	71%	72%
10	31%	71%	74%
15	38%	94%	94%

best to perform retrieval using 15 basis functions and two-second length of query sample. From Table 2, we can also see the same phenomena as Table 1 except for the lower recall rate.

Besides, by examining those missing in the experiments based on human judgement as the ground truth, we found two major factors. First, for the first experiment, we find that some errors occur in those searched clips containing a transition, which is made because we simply segment an audio object into several one-second clips uniformly against predividing the audio object into sequences of audio phrases. As a matter of fact, these kind of errors can be reduced by increasing the length of query sequence (i.e. clip number) to get more related information or performing the predivision for the audio phrases. Secondly, we find that some errors occur since the refrains of some songs are performed at different tempo. From these tables, we can see that the proposed retrieval approach for music data can achieve over 96% accuracy. The experiments are carried out on a Pentium II 400 PC/Windows 2000. One hundred and fifty queries can be processed in less than fourteen seconds for 10,000 phrases. In order to make a comparison, we would also like to cite the efficiency of the existing system described in Refs. 26 and 27, which also uses a similar database to ours. The authors reported that their accuracy rates are more than 90%.

4. Conclusions

Digital audio signals, especially for music are an important type of media. However, few works have been focused on the music databases. In this paper, we have presented a new method for content-based music retrieval to retrieve perceptually similar music pieces in audio documents. In the proposed method, based on the Gabor wavelet filters, the extracted perceptual features are general enough to meet the human auditory system. An accurate retrieval rate higher than 96% was achieved. Furthermore, the complexity is low due to the easy computing of audio features, and this makes online processing possible.

There are several related tasks to be conducted in the future. First, we will work on the other type of audio source such as sound effects and compression domain. Second, we will work on developing an automatic segmentation technique to divide the musical objects into sequences of phrase.

Acknowledgment

The authors would like to thank the anonymous reviewers for their many valuable suggestions which have greatly improved the presentation of the paper.

References

1. S.-T. Bow, *Pattern Recognition and Image Preprocessing* (Marcel Dekker, 1992).
2. J. Foote, Content-based retrieval of music and audio, in *Proc. SPIE, Multimedia Storage and Archiving Systems II*, **3229** (1997), pp. 138–147.
3. J. Foote, An overview of audio information retrieval, *ACM Multimed. Syst.* **7**(1) (1999) 2–11.

4. D. Gabor, Theory of communication, *J. Inst. Elect. Eng.* **93** (1946) 429–439.
5. A. Ghias, J. Logan, D. Chamberlin and B. Smith, Query by humming: Musical information retrieval in an audio database, in *Proc. Int. Conf. ACM Multimedia* (1995), pp. 213–236.
6. L. Guojun and T. Hankinson, A technique towards automatic audio classification and retrieval, in *Proc. Int. Conf. Signal Processing'98* **2** (1998), pp. 1142–1145.
7. ISO/IEC JTC1/SC29/WG11 Moving Pictures Expert Group, "Information technology-multimedia content description interface — Part 4: Audio. Comittee Draft 15938-4," ISO/IEC, 2000.
8. D. Kimber and L. D. Wilcox, Acoustic segmentation for audio browsers, in *Proc. Interface Conf.*, Sydney, Australia (July 1996).
9. D. Li, I. K. Sethi, N. Dimitrova and T. McGee, Classification of general audio data for content-based retrieval, *Patt. Recogn. Lett.* **22**(5) (2001) 533–544.
10. B. S. Manjunath and W. Y. Ma, Texture features for browsing and retrieval of image data, *IEEE Trans. Patt. Anal. Mach. Intell.* **18**(8) (1992) 173–188.
11. B. S. Manjunath, P. Salembier and T. Sikora, *Introduction to MPEG-7: Multimedia Content Description Interface* (John Wiley, 2002).
12. C. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing* (MIT Press, Cambridge, MA, 1999).
13. K. Martin, E. Scheirer and B. Vercoe, Musical content analysis through models of audition, in *Proc. ACM Multimedia Workshop on Content-Based Processing of Music*, Bristol, UK (1998).
14. MPEG Requirements Group, Description of MPEG-7 content set, Doc. ISO/MPEG N2467, MPEG Atlantic City Meeting (October 1998).
15. S. Pfeiffer, S. Fischer and W. Effelsberg, Automatic audio content analysis, in *Proc. ACM Multimedia'96*, Boston, MA (April 1996), pp. 21–30.
16. S. Qian and D. Chen, *Joint Time-Frequency Analysis Methods and Applications* (Prentice-Hall, Upper Saddle River, NJ, 1966).
17. F. D. Rosenthal, *Computational Auditory Scene Analysis* (Lawrence Erlbaum Associates, Inc., 1998).
18. G. Smith, H. Murase and H. Kashino, Quick audio retrieval using active search, in *Proc. Int. Conf. Acoustics, Speech, Signal Processing'98*, Seattle, WA (May 1998), pp. 3777–3780.
19. J. M. Smith and X. Serra, An analysis/resynthesis program for non-harmonic sounds based on a sinusoidal representation, in *Proc. ICMC 87*, Ann Arbor, Michigan (1987), 290 pp.
20. C. Spevak and E. Favreau, Soundspotter — a prototype system for contest-based audio retrieval, in *Proc. Int. Conf. Digital Audio Effects* (September 2002), pp. 27–32.
21. G. Tzanetakis, *Manipulation, Analysis and Retrieval System for Audio Signals*, Ph.D. thesis, Princeton University (2002).
22. G. Tzanetakis and P. Cook, Audio information retrieval (AIR) tools, in *Proc. Int. Symposium on Music Information Retrieval (ISMIR)* (2000).
23. A. J. Willis and L. Myers, A cost-effective fingerprint recognition system for use with low-quality prints and damaged fingertips, *Patt. Recogn.* **34**(2) (2001) 255–270.
24. E. Wold, T. Blum, D. Keislar and J. Wheaton, Content-based classification, search, and retrieval of audio, *IEEE Multimed.* **3**(3) (1996) 27–36.
25. L. Wyse and S. Smoliar, Toward content-based audio indexing and retrieval and a new speaker discrimination technique, in *Proc. ICJAI'95*, Singapore (December 1995).
26. C. Yang, MACS: music database retrieval based on spectral similarity, *IEEE Workshop on Applications of Signal Processing* (2001).

27. C. Yang, *Music Database Retrieval Based on Spectral Similarity*, Stanford University Database Group Technical Report 2001-14 (2001).
28. T. Zhang and C.-C. J. Kuo, Content-based classification and retrieval of audio, in *Proc. SPIE, Conf. Advanced Signal Processing Algorithm, Architectures, and Implementations VIII*, Vol. 3461, San Diego (July 1998).
29. T. Zhang and C.-C. J. Kuo, Hierarchical classification of audio data for archiving and retrieving, in *Proc. Int. Conf. Acoustics, Speech, Signal Processing'99* **6** (1999), pp. 3001–3004.
30. E. Zwicker and H. Fastl, *Psychoacoustics, Facts and Models* (Springer, 1990).

Ruei-Shiang Lin received the B.S. and M.S. degrees in electrical engineering from Tamkang University, Taiwan, in 1996 and Tatung University, Taiwan, in 1998, respectively. He is currently a Ph.D. student in the Department of Computer and Information Science at National Chiao Tung University, Taiwan.

His research interests include image processing, pattern recognition and audio analysis.

Ling-Hwei Chen received the B.S. degree in mathematics and the M.S. degree in applied mathematics from National Tsing Hua University, Hsinchu, Taiwan in 1975 and 1977, respectively, and the Ph.D. in computer engineering from National Chiao Tung University, Hsinchu, Taiwan in 1987.

From August 1977 to April 1979, she worked as a research assistant in the Chung-Shan Institute of Science and Technology, Taoyan, Taiwan, from May 1979 to February 1981, she worked as a research associate in the Electronic Research and Service Organization, Industry Technology Research Institute, Hsinchu, Taiwan. From March 1981 to August 1983, she worked as an engineer in the Institute of Information Industry, Taipei, Taiwan. She is now a Professor in the Department of Computer and Information Science at the National Chiao Tung University.

Her current research interests include image processing, pattern recognition, video/image compression, and multimedia steganography.