International Journal of Pattern Recognition
and Artificial Intelligence
Vol. 19, No. 6 (2005) 807–822

World Scientific
www.worldscientific.com

A NEW APPROACH FOR AUDIO CLASSIFICATION AND SEGMENTATION USING GABOR WAVELETS AND FISHER LINEAR DISCRIMINATOR*

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@cis.nctu.edu.tw*

Rapid increase in the amount of audio data demands an efficient method to automatically segment or classify audio stream based on its content. In this paper, based on the Gabor wavelet features, an audio classification and segmentation method is proposed. This method will first divide an audio stream into clips, each of which contains one-second audio information. Then, each clip is classified as one of two classes or five classes. Two classes contain speech and music; pure speech, pure music, song, speech with music background, and speech with environmental noise background are for five classes. Finally, a merge technique is provided to do segmentation.

In order to make the proposed method robust for a variety of audio sources, we use Fisher Linear Discriminator to obtain features with the highest discriminative ability. Experimental results show that the proposed method can achieve over 98% accuracy rate for speech and music discrimination, and more than 95% for a five-way discrimination. By checking the class types of adjacent clips, we can also identify more than 95% audio scene breaks in audio sequence.

Keywords: Audio classification and segmentation; spectrogram; audio content-based retrieval; Fisher Linear discriminator; Gabor wavelets.

1. Introduction

In recent years, audio, as an important and integral part of many multimedia applications, has gained more and more attention. Rapid increase in the amount of audio data demands an efficient method to automatically segment or classify audio stream based on its content. Many studies on audio content analysis[2,4,5,7–9,12,15–17,20–22] have been proposed.

A speech/music discriminator was provided in Ref. 17, based on thirteen features including cepstral coefficients, four multidimensional classification frameworks are compared to achieve better performance. The approach presented by Saunders[16]

*This research was supported in part by the National Science Council of R.O.C. under contract NSC-90-2213-E-009-127.
†Author for correspondence.

takes a simple feature space, it is performed by exploiting lopsideness of the distribution of zero-crossing rate, where speech signals show a marked rise that is not common for music signals. In general, for speech and music, it is not hard to reach a relatively high level of discrimination accuracy since, different properties exist in both time and frequency domains.

Besides speech and music, it is necessary to take other kinds of sounds into consideration in many applications. The classifier proposed by Wyse and Smoliar[20] classifies audio signals into "music", "speech", and "others". It was developed for the parsing of news stories. In Ref. 8, audio signals are classified into speech, silence, laughter, and nonspeech sounds for the purpose of segmenting discussion recordings in meetings. However, the accuracy of the segmentation results using this method varies considerably for different types of recording. Besides the commonly studied audio types such as speech and music, research in Refs. 9, 21 and 22 has taken into account hybrid-type sounds, e.g. speech signal with music background and the singing of a person, which contain more than one basic audio type and usually appear in documentaries or commercials. In Ref. 9, 143 features are first studied for their discrimination capability. Then, the cepstral-based features such as Mel-frequency cepstral coefficients (MFCC), linear prediction coefficients (LPC), etc. are selected to classify audio signals. Zhang and Kuo[22] extracted some audio features including the short-time fundamental frequency and the spectral tracks by detecting the peaks from the spectrum. The spectrum is generated by autoregressive model (AR model) coefficients, which are estimated from the autocorrelation of audio signals. Then, the rule-based procedure, which uses many threshold values, is applied to classify audio signals into speech, music, song, speech with music background, etc. Accuracy of the above 90% is reported. However, this method is complex and time-consuming due to the computation of autocorrelation function. Besides, the thresholds used in this approach are empirical, they are improper when the source of audio signals is changed.

In this paper, we will provide two classifiers, one is for speech and music (called two-way); the other is for five classes (called five-way) that are pure speech, music, song, speech with music background, and speech with environmental noise background. Based on the classification results, we will propose a merging algorithm to divide an audio stream into some segments of different classes.

One basic issue for content-based classification of audio sound is feature selection. The selected features should be able to represent the most significant properties of audio sounds, and they are also robust under various circumstances and general enough to describe various sound classes. The issue in the proposed method is addressed in the following: first, some perceptual features based on the Gabor wavelet filters[6,10] are extracted as initial features, then Fisher Linear Discriminator (FLD)[3] is applied to these initial features to explore the features with the highest discriminative ability.

Note that FLD is a tool for multigroup data classification and dimensionality reduction. It maximizes the ratio of between-class variance to within-class variance

in any particular data set to guarantee maximal separability. Experimental results show that the proposed method can achieve an accuracy rate of discrimination over 98% for a two-way speech/music discriminator, and more than 95% for a five-way classifier which uses the same database as that used in the two-way discrimination. Based on the classification result, we can also identify scene breaks in audio sequence quite accurately. Experimental results show that our method can detect more than 95% of audio type changes. 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 five phases: time-frequency distribution (TFD) generation, initial feature extraction, feature selection, classification and segmentation. First, the input audio is transformed to a spectrogram, which will meet the ear-hearing system. 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.

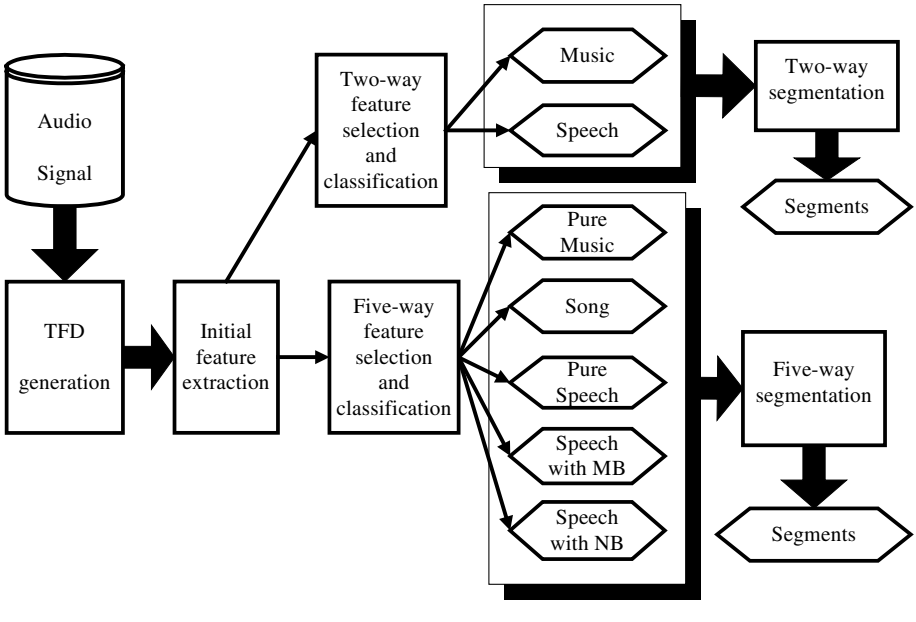


Fig. 1. Block diagram of the proposed method, where "MB" and "NB" are the abbreviations for "music background" and "noise background", respectively.

Third, based on the extracted initial features, the Fisher Linear Discriminator (FLD) is used to select the features with the best discriminative ability and also to reduce feature dimension. Fourth, based on the selected features, a classification method is then provided to classify each clip. Finally, based on the classified clips, a segmentation technique is presented to identify scene breaks in each audio stream. 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).[13] Conventionally, the Short Time Fourier Transform (STFT) is applied to construct a spectrogram and the TFD is sampled uniformly in time and frequency. However, it is not suitable for the auditory model because the frequency resolution within the human psycho-acoustic system is not constant but varies with frequency.[23]

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.[23] Given the sampling frequency (Fs) of 441,00 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 each of 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.

There are three parts in the TFD generation. In the first part, the N-point STFT is applied to the original audio signal $P_1(t)$ to obtain a spectrogram $S_1(x, y)$. In the second part, $P_1(t)$ is downsampled to half-size to obtain signal $P_2(t)$ and the N-point STFT is applied to $P_2(t)$ to obtain a spectrogram $S_2(x, y)$. In the third part, $P_1(t)$ is downsampled to quarter-size to obtain signal $P_3(t)$ and the N-point STFT 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 original signal to prevent the aliasing, and the window size for STFT is 512 (i.e. $N = 512$) in this

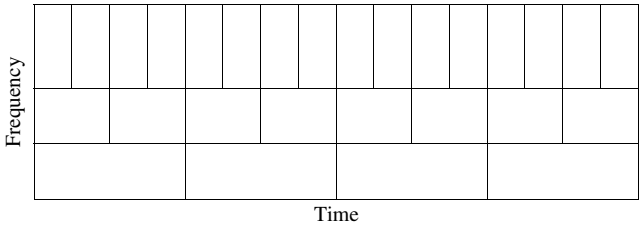


Fig. 2. An example of tiling in the time-frequency plane.

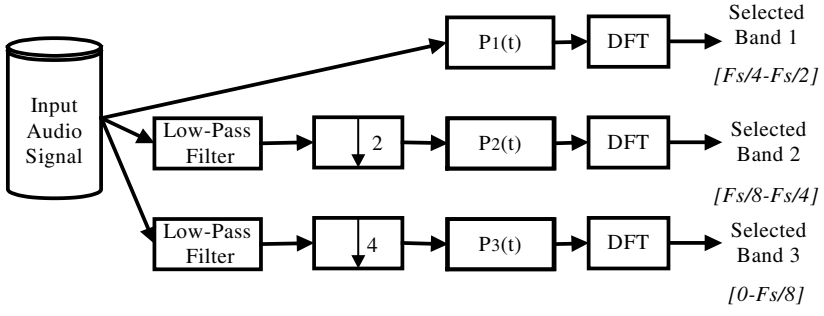


Fig. 3. A schematic diagram of the TFD generation details.

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

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

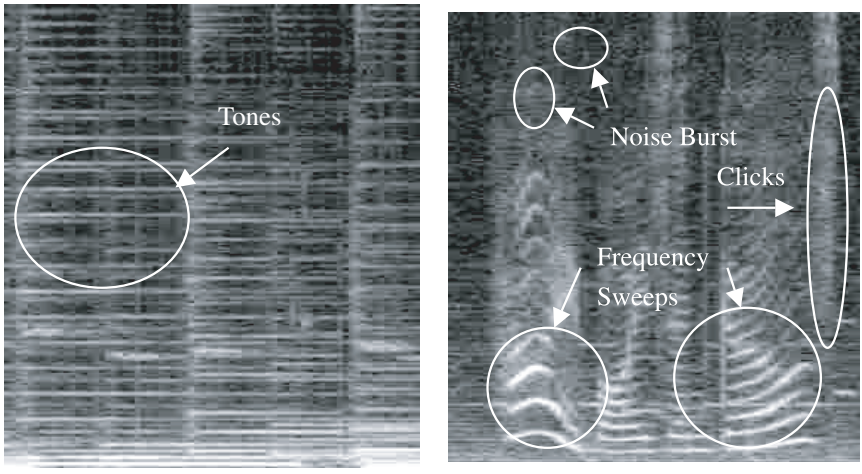
Finally, based on $S_1(x, y)$, $S_2(x, y)$, and $S_3(x, y)$, a spectrogram $I(x, y)$ is obtained according to the following equation:

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

where N_f is the frame number of $P_1(t)$. From Eq. (3), we can see that in $I(x, y)$, the 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.[14] All curve-like patterns are called tracks. Figure 4(a) shows that for a music 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), and frequency sweeps in a song spectrogram. Thus, if we can extract some features from a spectrogram to represent these patterns, the



(a) (b)

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

classification should be easy. Smith and Serra[18] 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,[13] in this paper, we will use the Gabor wavelet functions to extract some initial features to represent those patterns. The details 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[6]:

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

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

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

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 frequency, respectively. In this paper, we set $\omega_l = 3/64$, $\omega_h = 3/4$, $K = 6$ and $S = 7$.

2.2.2. *Feature estimation and representation*

To extract the audio features, each Gabor wavelet filter, $g_{mn}(x, y)$, is first applied to the spectrogram $I(x, y)$ to get a filtered spectrogram, $W_{mn}(x, y)$, as

$$W_{mn}(x,y) = \int I(x - x_1, y - y_1)\, g_{mn} * (x_1, y_1) dx_1 dy_1, \tag{10}$$

where * indicates the complex conjugate. The above filtering process is executed by *FFT* (Fast Fourier Transform). That is

$$W_{mn}(x,y) = F^{-1}\{F\{g_{mn}(x,y)\} \cdot F\{I(x,y)\}\}. \tag{11}$$

Since peripheral frequency analysis in the ear system roughly follows a logarithmic axis, in order to keep this way, the entire frequency band $[0, Fs/2]$ is divided into six subbands of unequal width: F1 = $[0, Fs/64]$, F2 = $[Fs/64, Fs/32]$, F3 = $[Fs/32, Fs/16]$, F4 = $[Fs/16, Fs/8]$, F5 = $[Fs/8, Fs/4]$, and F6 = $[Fs/4, Fs/2]$. In our experiments, high frequency components above $Fs/4$ (i.e. subband $[Fs/4, Fs/2]$) are discarded to avoid the influence of noise. Then, for each interested subband F_i, the directional histogram, $H_i(m, n)$, is defined to be

$$H_i(m,n) = \frac{N_i(m,n)}{\sum_{n=0}^{5} N_i(m,n)}, \quad i = 0, \ldots, 4, \tag{12}$$

$$W_{mn}^i(x,y) = \begin{cases} 1, & \text{if } W_{mn}(x,y) > T_m \quad \text{and} \quad y \in F_i \\ 0, & \text{otherwise,} \end{cases} \tag{13}$$

$$N_i(m,n) = \sum_x \sum_y W_{mn}^i(x,y), \tag{14}$$

where $m = 0, \ldots, 6$ and $n = 0, \ldots, 5$. Note that $N_i(m, n)$ is the number of pixels in the filtered spectrogram $W_{mn}(x, y)$ at subband F_i, scale m and direction n with value larger than threshold T_m. T_m is set as

$$T_m = \mu_m + \sigma_m, \tag{15}$$

where $\mu_m = \sum_{n=0}^{5} \sum_x \sum_y W_{mn}(x, y)/N_m$, $\sigma_m = (\sum_{n=0}^{5} \sum_x \sum_y (W_{mn}(x, y) - \mu_m)^2 / N_m)^{\frac{1}{2}}$, and N_m is the number of pixels over all the six filtered spectrogram $W_{mn}(x, y)$ with scale m.

An initial feature vector, f, is now constructed using $H_i(m, n)$ as feature components. Recall that in our experiments, we use seven scales ($S = 7$), six orientations ($K = 6$) and five subbands, this will result in a $7 \times 6 \times 5$-dimensional initial feature vector

$$f = [H_0(0, 0),\ H_0(0, 1), \ldots, H_4(6, 5)]^T. \tag{16}$$

2.3. *Feature selection and audio classification*

The initial features are not used directly for classification since some features give poor separability among different classes and inclusion of these features will lower down classification performance. In addition, some features are highly correlated so that redundancy will be introduced. To remove these disadvantages, in this paper, the Fisher Linear Discriminator (FLD) is applied to the initial features to find those uncorrected features with the highest separability. Before describing FLD, two matrices, between-class scatter and within-class scatter, will first be introduced. The within-class scatter matrix measures the amount of scatter between items in the same class and the between-class scatter matrix measures the amount of scatter between classes.

For the ith class, the within-class scatter matrix S_w^i is defined as

$$S_w^i = \sum_{x_k^i \in X_i} (x_k^i - \mu_i)(x_k^i - \mu_i)^T, \tag{17}$$

the total within-class scatter matrix S_w is defined as

$$S_w = \sum_{i=1}^{C} S_w^i, \tag{18}$$

and the between-class scatter matrix S_b is defined as

$$S_b = \sum_{i=1}^{C} N_i(\mu_i - \mu)(\mu_i - \mu)^T, \tag{19}$$

where μ_i is the mean of class X_i, N_i is the number of samples in class X_i, x_k^i is the kth sample in X_i, and C is the number of classes.

In FLD, a matrix $V_{\text{opt}} = \{v_1, v_2, \ldots, v_{C-1}\}$ is first chosen, it satisfies the following equation:

$$V_{\text{opt}} = \arg\max_{V} \left| \frac{V^T S_b V}{V^T S_w V} \right|. \tag{20}$$

In fact, $\{v_1, v_2, \ldots, v_{C-1}\}$ is the set of generalized eigenvectors of S_b and S_w corresponding to the $C-1$ largest generalized eigenvalues $\{\lambda_i | i = 1, 2, \ldots, C-1\}$,[3] i.e.

$$S_b v_i = \lambda_i S_w v_i. \tag{21}$$

Note that in this paper, two classes and five classes (i.e. $C = 2$ and $C = 5$) are used and one-second audio clip is taken as the basic classification unit.

Based on V_{opt}, the initial feature vector for each one-second audio clip in the training data and testing data is projected to the space generated by V_{opt} to get a new feature vector f' with dimension $C-1$. f' is then used to stand for the audio clip. Before classification, it is important to give a good similarity measure. In our experiments, the Euclidean distance worked better than others (e.g. Mahalanobis, covariance, etc.). For each test sample, x_j with feature vector f'_j, the Euclidean distance between the test sample and the class center of each class in the space generated by V_{opt} is evaluated. Then the sample is assigned to the class with minimum distance. That is, x_j is assigned as class C'_j according to the following criterion:

$$C'_j = \arg\min_{i} \left\| f'_j - \mu'_i \right\|, \quad i = 1, 2, \ldots, C, \tag{22}$$

where μ'_i is the mean vector of the projected vectors of all training samples in class i. Figure 5 shows an example of using a two-way speech/music discriminator. In the figure, "x" stands for the projected result of a music signal, "o" stands for the projected result of a speech signal. From this figure, we can see that through FLD, music and speech samples can be easily separated. Figure 6 outlines the process of feature selection and classification.

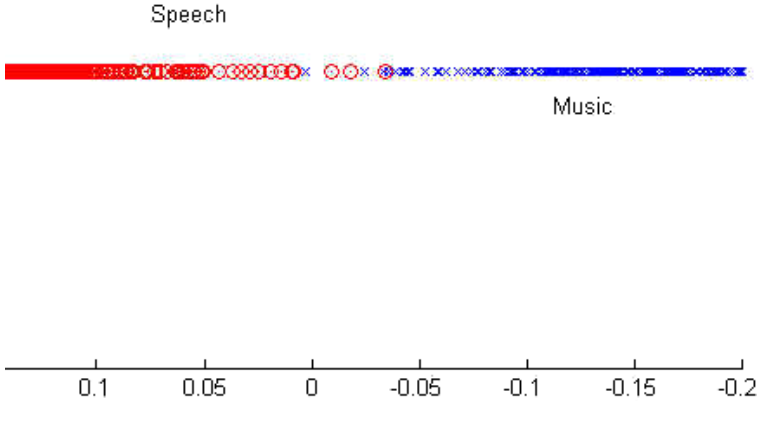


Fig. 5. An example of using FLD for two-way speech/music discriminator.

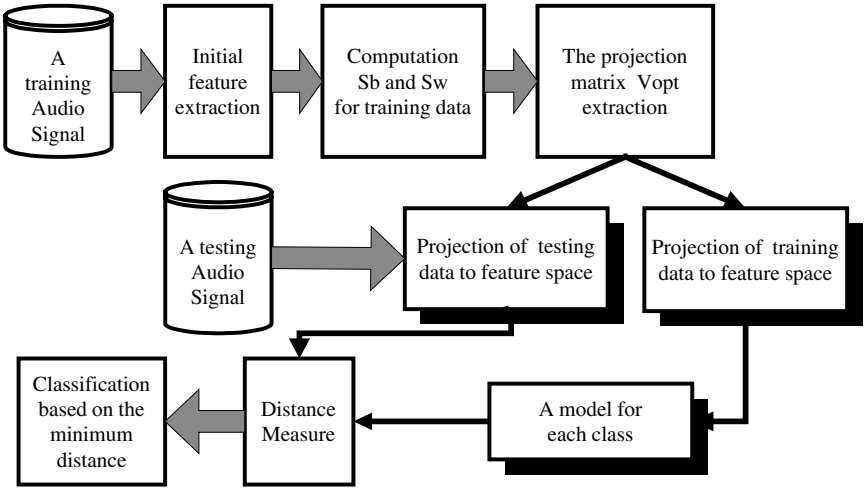


Fig. 6. A block diagram of feature selection and classification using FLD.

Two problems arise when using Fisher discriminator. First, the matrices needed for computation are very large. Second, since we may have fewer training samples than the number of features in each sample, the data matrix is rank deficient. To avoid the problems described above, it is possible to solve the eigenvectors and eigenvalues of a rank deficient matrix by using a generalized singular value decomposition routine. One simple and speedup solution[1] is taken in this paper.

2.4. *Segmentation*

The segmentation is to divide an audio sequence into semantic scenes called "audio scene" and to index them as different audio classes. Due to some classification errors, a reassigning algorithm is first provided to rectify these classification errors. For example, if we detect a pattern like speech-music-speech, and the music subpattern lasts a very short time, we can conclude that the music subpattern should be speech. First, for each one-second audio clip, the similarity measure between the audio clip and the center of its class is defined as

$$\text{Similarity} = 1 - \frac{\text{dist}_{\min}}{\sum_{j=1}^{5} \text{dist}_j}, \quad \text{dist}_{\min} = \min_j \text{dist}_j, \tag{23}$$

where dist_j is the Euclidean distance between the clip and the jth class center in the feature space. If the similarity measure is less than 0.9, mark the clip as ambiguous. Note that ambiguous clips often arise in transition periods. For example, if a transition happens when speech stops and music starts, then each clip in the transition will contain both speech and music information. Then, each ambiguous clip will be reassigned as the class of the nearest unambiguous clip. After the reassignment is completed, all neighboring clips with the same class are merged into a segment. Finally, for each audio segment, the length is evaluated. If the length is shorter

than the threshold T ($T = 3$ second), each clip in the segment is reassigned as the class of one of its two neighboring audio segments with the least Euclidean distance between the clip and the center of class of the selected neighboring segment.

3. Experimental Results

3.1. *Audio database*

In order to do comparison, we have collected a set of 700 generic audio pieces (with duration from several seconds to no more than one minute) of different types of sound according to the collection rule described in Ref. 22 as the testing database. Care was taken to obtain a wide variation in each category, and some clips are taken from MPEG-7 content set.[11] The database contains 100 pieces of classical music played with various instruments, 100 other music pieces of different styles (jazz, blues, light music, etc.), 200 pieces of pure speech in different languages (English, Chinese, Japanese, etc.), 200 pieces of song sung by male, female, or children, 50 pieces of speech with background music (e.g. commercials, documentaries, etc.), and 50 pieces of speech with environmental noise (e.g. sport broadcast, news interview, etc.). These shorter 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 classification performance. Note that we take one-second audio signal as a test unit.

We have also collected a set of 15 longer audio pieces recorded from movies, radio or video programs. These pieces last from several minutes to an hour and contain various types of audio. They are used to test the performance for audio segmentation.

3.2. *Classification and segmentation results*

3.2.1. *Audio classification results*

In order to examine the robust use for a variety of the audio source and the accuracy for audio classification, we present two experiments. One is two-way discrimination and the other is five-way discrimination. Concerning the two-way discrimination, we try to classify the audio set into two categories: music and speech. As for the five-way discrimination, the audio set will be classified into five categories: pure speech, pure music, song, speech with music background, and speech with environmental noise background.

Tables 1 and 2 show the results of the classification. From these tables, we can see that the proposed classification approach for generic audio data can achieve an

Table 1. Two-way classification results.

Audio Type	Number	Correct Rate
Speech	300	98.17%
Music	400	98.79%

Table 2. Five-way classification results.

		Discrimination Results				
Audio Type	Number	Pure Music	Song	Pure Speech	Speech with MB	Speech with NB
Pure Music	200	94.67%	3.21%	1.05%	1.07%	0%
Song	200	0.8%	96.43%	0%	1.97%	0.8%
Pure Speech	200	0%	0.14%	98.40%	0.11%	1.35%
Speech with MB	50	1.01%	4.2%	3.10%	89.62%	2.07%
Speech with NB	50	0.15%	0.71%	1.28%	0.63%	97.23%

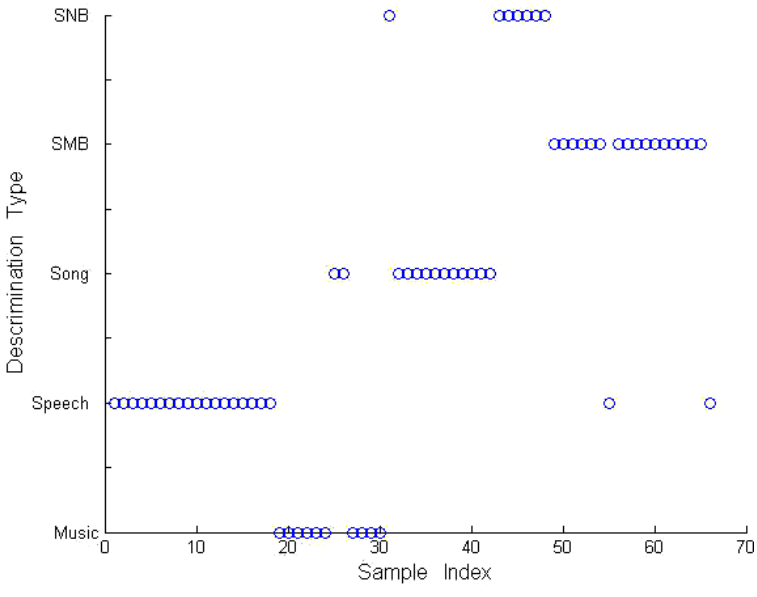
over 98% accuracy rate for the speech/music discrimination, and more than 95% for the five-way classification. Both classifiers use the same testing database. It is worth mentioning that the training is done using 50% of randomly selected samples in each audio type, and the test is operated on the remaining 50%. By changing training set several times and evaluating the classification rates, we find that the performance is stable and independent on the particular test and training sets. The experiments are carried out on a Pentium II 400 PC/Windows 2000 with less than one-eleventh of the time required to play the audio clip.

In our experiments, there are several misclassifications. From Table 2, we can see that most errors occur in the speech with music background category. This is because the music or speech component is weak. In order to do a comparison, we would also like to cite the efficiency of the existing system described in Ref. 22 which also includes the five audio classes considered in our method and use databases similar to ours. The authors of Ref. 22 report that less than one eighth of the time required to play the audio clip are needed to process an audio clip. They also report that their accuracy rates are more than 90%.

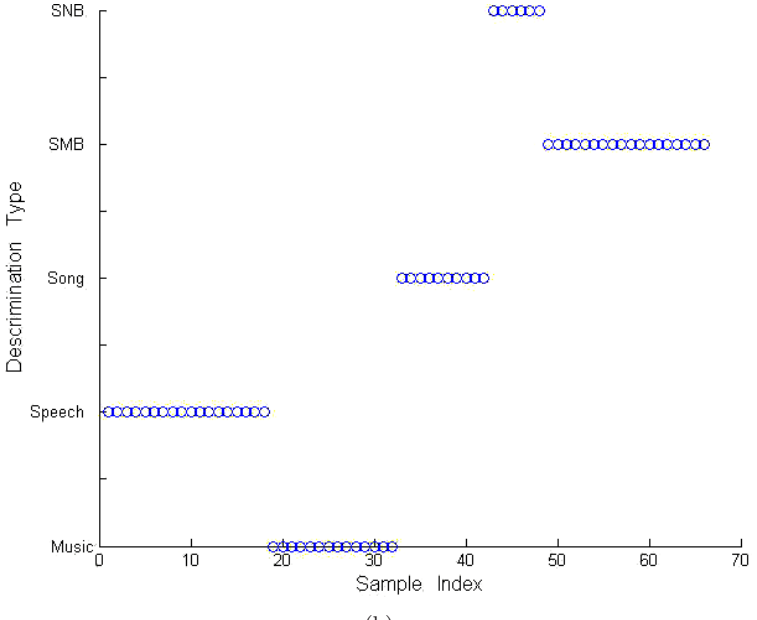
3.2.2. *Audio segmentation results*

We tested our segmentation procedure with audio pieces recorded from radio, movies and video programs. We made a demonstration program for online audio segmentation and indexing as shown in Fig. 7. Figure 7(a) shows the classification result for a 66 second audio piece recorded from MPEG-7 data set CD19 that is a Spanish cartoon video called "Don Quijote de la Mancha". Figure 7(b) shows the result of applying the segmentation method to Fig. 7(a). Besides the above example, we have also performed experiments on other audio pieces.

Listed in Table 3 is the result of the audio segmentation, where miss-rate and over-rate are defined as the ratio between the number of miss-segmented ones and the actual number of segments, and the ratio between the number of over-segmented ones and the actual number of segments in audio streams, respectively. Besides, error rate is defined as the ratio between the number of segments indexed in errors and the actual number of segments in audio stream.



(a)



(b)

Fig. 7. Demonstration of audio segmentation and indexing, where "SMB" and "SNB" are the abbreviations for "speech with music background" and "speech with noise background", respectively. (a) Original result. (b) Final results after applying the segmentation algorithm to (a).

Table 3. Segmentation results.

	Without Using Reassignment	Using Reassignment
Miss-Rate	0%	1.1%
Over-Rate	5.2%	1.8%
Error-Rate	2.5%	1.3%

The first column shows the segmentation result without applying the reassignment process to the classification result, and the second column shows the segmentation result using the reassignment process. Experiments have shown that the proposed scheme achieves satisfactory segmentation and indexing. Using human judgement as the ground truth, our method can detect more than 95% of audio type changes.

4. Conclusions

In this paper, we have presented a new method for the automatic classification and segmentation of generic audio data. An accurate classification rate higher than 95% was achieved. The proposed scheme can treat a wide range of audio types. Furthermore, the complexity is low due to the easy computing of audio features, and this makes online processing possible. The experimental results indicate that the extracted audio features are quite robust.

Besides the general audio types such as music and speech tested in existing work, we have taken into account other different types of sounds including hybrid-type sounds (e.g. speech with music background, speech with environmental noise background, and song). While current existing approaches for audio content analysis are normally developed for specific scenarios, the proposed method is generic and model free. Thus, it can be widely applied to many applications.

References

1. N. P. Belhumeur and D. J. Kriegman, Eigenfaces vs. fishfaces: recognition using class specific linear projection, *IEEE Trans. Patt. Anal. Mach. Intell.* **19**(7) (1997) 711–720.
2. J. S. Boreczky and L. D. Wilcox, A hidden Markov model framework for video segmentation using audio and image features, in *Proc. Int. Conf. Acoustics, Speech, Signal Processing'98* (May 1998), pp. 3741–3744.
3. S.-T. Bow, *Pattern Recognition and Image Preprocessing* (Marcel Dekker, 1992).
4. J. Foote, An overview of audio information retrieval, *ACM Multimed. Syst.* **7**(1) (1999) 2–11.
5. I. Fujinaga, Machine recognition of timbre using steady-state tone of acoustic instruments, in *Proc. ICMC 98*, Ann Arbor, Michigan (1998), pp. 207–210.
6. D. Gabor, Theory of communication, *J. Instit. Electr. Eng.* **93** (1946) 429–439.
7. L. Guojun and T. Hankinson, A technique towards automatic audio classification and retrieval, in *Proc. Int. Conf. Signal Processing'98*, Vol. 2 (1998), pp. 1142–1145.
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. MPEG Requirements Group, Description of MPEG-7 content set, Doc. ISO/MPEG N2467, MPEG Atlantic City Meeting (October 1998).
12. S. Pfeiffer, S. Fischer and W. Effelsberg, Automatic audio content analysis, in *Proc. ACM Multimedia'96*, Boston, MA (April 1996), pp. 21–30.
13. S. Qian and D. Chen, *Joint Time-Frequency Analysis Methods and Applications* (Prentice-Hall, Upper Saddle River, NJ, 1996).
14. F. D. Rosenthal, *Computational Auditory Scene Analysis* (Lawrence Erlbaum Associates, Inc., 1998).
15. S. Rossignol, X. Rodet *et al.*, Feature extraction and temporal segmentation of acoustic signals, in *Proc. ICMC 98*, Ann Arbor, Michigan (1998), pp. 199–202.
16. J. Saunders, Real-time discrimination of broadcast speech/music, in *Proc. Int. Conf. Acoustics, Speech, Signal Processing'96*, Vol. 2, Atlanta, GA (May 1996), pp. 993–996.
17. E. Scherier and M. Slaney, Construction and evaluation of a robust multifeature speech/music discriminator, in *Proc. Int. Conf. Acoustics, Speech, Signal Processing'97*, Munich, Germany, April 1997, pp. 1331–1334.
18. J. 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.
19. E. Wold, T. Blum, D. Keislar and J. Wheaton, Content-based classification, search, and retrieval of audio, *IEEE Multimed.* **3**(3) (1996) 27–36.
20. 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).
21. 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*, Vol. 6 (1999), pp. 3001–3004.
22. T. Zhang and C.-C. J. Kuo, Audio content analysis for online audiovisual data segmentation and classification, *IEEE Trans. Speech Audio Process.* **9**(4) (2001) 441–457.
23. 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, and the Ph.D. from National Chiao Tung University, Taiwan, in 2004. In 2005, he joined Leadtek Research Inc., Taiwan, where he is currently a Senior Engineer involved in the development of video codecs.

His research interests include image and video 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.