A Robust Audio Classification and Segmentation Method

Lie Lu, Hao Jiang and HongJiang Zhang
Microsoft research, China
5F, Beijing Sigma Center
No.49 Zhichun Road, Beijing 100080, PRC
Phone:+86-10-62617711
{i-lielu, hjzhang}@microsoft.com

ABSTRACT
In this paper, we present a robust algorithm for audio classification that is capable of segmenting and classifying an audio stream into speech, music, environment sound and silence. Audio classification is processed in two steps, which makes it suitable for different applications. The first step of the classification is speech and non-speech discrimination. In this step, a novel algorithm based on KNN and *LSP* VQ is presented. The second step further divides non-speech class into music, environment sounds and silence with a rule based classification scheme. Some new features such as the *noise frame ratio* and *band periodicity* are introduced and discussed in detail. Our experiments in the context of video structure parsing have shown the algorithms produce very satisfactory results.

Keywords
Audio content analysis, audio classification and segmentation

1. INTRODUCTION
Rapid increase in the amount of audio data demands for an efficient method to automatically segment or classify audio stream based on its content. Such a method is helpful not only in audio retrieval [1][2], but also in video structure extraction.

In general, audio content analysis in video parsing can be considered in two directions [12][13][14]. One is to discriminate audio streams into different classes such as speech, music, environment sound and silence, the other is to classify audio streams into segments of different speakers. In this paper, our research work of the first direction will be presented.

There have been many studies on audio content analysis, using different features and different methods. In spite of many research efforts, high accuracy audio classification is only achieved for the simple cases such as speech/music discrimination. Pfeiffer et al [3], presented a theoretic framework and application of automatic audio content analysis using some perceptual features. Saunders [4], presented a speech/music classifier based on simple features such as zero crossing rate and short time energy for radio broadcast. When a window size of 2.4s was used, the reported accuracy rate would be 98%. Scheirer et al [5] introduced many more features into audio classification and performed experiments with different classification models including GMM (Gaussian Mixture Model), BP-ANN (Back Propagation Artificial Neural Network) and KNN (K-Nearest Neighbor). When using window of the same size (2.4s), the reported error rate would be 1.4%. However, it is found that such simple features based methods cannot work well when smaller window is used or more audio classes such as environment sounds are taken into consideration.

Many other works have been done to enhance audio classification algorithms. In [6], audio recordings are classified into speech, silence, laughter and non-speech sounds, in order to segment discussion recordings in meetings. The accuracy of the segmentation resulted using his method varies considerably for different types of recording. In the work by Zhang and Kuo [7], pitch tracking methods are introduced to discriminate audio recordings into more classes, such as songs, speeches over music, with a heuristic-based model. Accuracy of above 90% is reported. Srinivasan et al [12], try to detect and classify audio that consists of mixed classes, such as combinations of speech and music together with background sound. The accuracy of classification is over 80%.

In this paper, a high accuracy algorithm of audio classification and segmentation for video structure paring is presented. We plan to discriminate speech, music, environment sound and silence in one-second window. They are the basic sets needed in video structure parsing. In order to classify these four audio classes more accurately, new feature, such as *band periodicity*, is proposed and discussed in detail.

The rest of the paper is organized as follows. Audio features are discussed in detail in Section 2. The classification and segmentation scheme is presented in Section 3. In Section 4, experiments and the evaluations of the proposed algorithms are given.

2. FEATURE ANALYSIS
In order to get high accuracy for classification and segmentation, we should select good features that can capture the temporal and spectral structures of audio. Grounded on the work in [5], we select following features: *high zero-crossing rate ratio* (*HZCRR*), *low short-time energy ratio* (*LSTE*R), and *spectrum flux* (*SF*). These parameters describe the variations of zero-crossing rate, short time energy and spectrum of an audio segment. We have

also introduced three new features: *LSP distance*, *band periodicity* (*BP*) and *noise frame ratio* (*NFR*), which are also very useful to classify speech, music and environment sound. Different features are used in different classifiers. All features are used to represent the characteristics of one-second audio segment.

2.1 High Zero-Crossing Rate Ratio

Zero-crossing rate (*ZCR*) is proved to be very useful in characterizing different audio signals. It was used in many previous speech/music classification algorithms. In our experiments, we found the variation of *ZCR* is more discriminative than the exact value of *ZCR*, so we use *high zero-crossing rate ratio* (*HZCRR*) as one feature in our algorithm.

HZCRR is defined as the ratio of the number of frames whose *ZCR* are above 1.5 fold average zero-crossing rate in one-second window, as following shows,

$$HZCRR = \frac{1}{2N}\sum_{n=0}^{N-1}[\mathrm{sgn}(ZCR(n) - 1.5avZCR) + 1] \quad (1.1)$$

$$avZCR = \frac{1}{N}\sum_{n=0}^{N-1} ZCR(n) \quad (1.2)$$

where n is the frame index, N is the total number of frames in a one-second window, *sgn*[.] is a sign function and $ZCR(n)$ is the zero-crossing rate at the nth frame, respectively.

In general, speech signals are composed of alternating voiced sounds and unvoiced sounds in the syllable rate, while music signals do not have this kind of structure. Hence, for speech signal, its variation of zero-crossing rates (or *HZCRR*) will be greater than that of music, as shown in Figure 1.

In Figure 1, the speech segment is from 0 to 200 seconds and its *HZCRR*s are around 0.15. The music segment is from 201 to 350 seconds, and its *HZCRR*s are around 0.05, while most of them are 0. Environment sound segment is from 351 to 450 seconds, and its *HZCRR*s vary dramatically. This is because there are many kinds of environment sound and their characteristics differ significantly. For example, for white noise, its *HZCRR* is low; but for the sound of drum, its *HZCRR* is high.

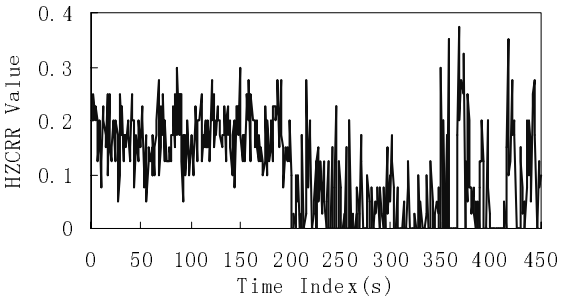

Figure 1. The high zero-crossing rate ratio curves (0-200s is speech; 201-350s is music and 351-450s is environment sound)

We also found that the *HZCRR* of speech is usually higher than that of music. In order to illustrate the discriminability of *HZCRR* more clearly, we extracted *HZCRR* for each one-second audio segment in our training database. From these data, we obtained the probability distribution curves of *HZCRR* for speech and music, illustrated in Figure 2.

Suppose we only use *HZCRR* to discriminate speech from music and use the cross-point of two curves as threshold, its error rate would be 19.36%.

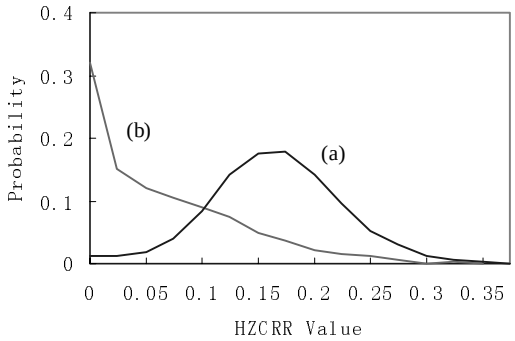

Figure 2. The probability distribution curves of *HZCRR*; (a) speech and (b) music

2.2 Low Short-Time Energy Ratio

Just as we have done to *ZCR*, we also selected the variation, not the exact value, of short-time energy as one component of our feature vector. Here, we use *low short-time energy ratio (LSTER)* to represent the variation of short-time energy (*STE*).

LSTER is defined as the ratio of the number of frames whose *STE* are less than 0.5 times of average short time energy in a one-second window, as the following,

$$LSTER = \frac{1}{2N}\sum_{n=0}^{N-1}[0.5avSTE - \mathrm{sgn}(STE(n)) + 1] \quad (2.1)$$

$$avSTE = \frac{1}{N}\sum_{n=0}^{N-1} STE(n) \quad (2.2)$$

where *N is* the total number of frames, *STE(n)* is the short time energy at the nth frame, and *avSTE* is the average *STE* in a one-second window.

LSTER is an effective feature, especially for discriminate speech and music signals. In general, there are more silence frames in speech, so the *LSTER* measure will be much higher for speech than that for music. This can be seen clearly from the probability distribution curves of *LSTER* for speech and music, which are illustrated in the Figure. 3. These curves are also obtained from our database by one-second windows. It is shown that *LSTER* value of speech is around 0.15 to 0.5, while most of music is less than 0.15. Therefore, *LSTER* is good discriminator for speech and music.

If we only use LSTER to discriminate speech from music and use the cross-point of two curves as threshold, its error rate would be only 8.27%. It can be easily calculated from the curves.

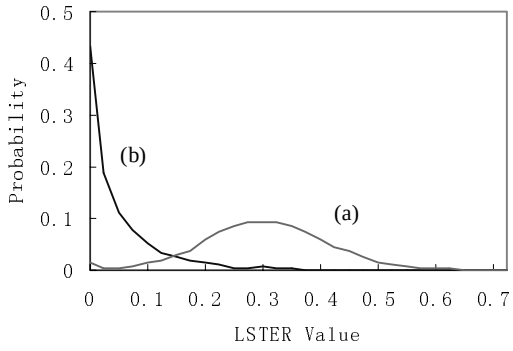

Figure 3. The probability distribution curves of *LSTER*; (a) speech and (b) music

2.3 Spectrum Flux

Spectrum Flux (SF) is defined as the average variation value of spectrum between the adjacent two frames in one second window,

$$SF = \frac{1}{(N-1)(K-1)} \sum_{n=1}^{N-1} \sum_{k=1}^{K-1} [\log(A(n,k)+\delta) - \log(A(n-1,k)+\delta]^2 \tag{3.1}$$

where *A(n, k)* is the Discrete Fourier Transform of the *n*th frame of input signal:

$$A(n,k) = \left| \sum_{m=-\infty}^{\infty} x(m) w(nL-m) e^{j\frac{2\pi}{L}km} \right| \tag{3.2}$$

and x(m) is the original audio data, w(m) the window function, *L* is the window length, *K* is the order of DFT, *N* is the total number of frames and δ a very small value to avoid calculation overflow.

In our experiments, we found that, in general, the *SF* values of speech are higher than those of music, and those of environment sound are the highest. Environment sounds also change more dramatically than the other two signals. Figure 4 shows an example of spectrum flux of speech, music and environment sound. The speech segment is from 0 to 200 seconds, the music segment is from 201 to 350 seconds and the environment sound is from 351 to 450 seconds.

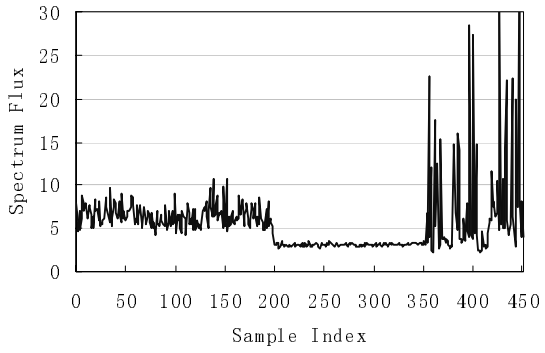

Figure 4. The spectrum flux curve (0-200s is speech; 201-350s is music and 351-450s is environment sound)

2.4 LSP Distance Measure

As the spectrum envelope parameter representation, linear predictive coefficient (LPC) is found effective for speech and non-speech discrimination. It is also found that linear prediction coefficients are more robust in the noisy environment. From the liner prediction coefficients, we can obtain the inverse filter:

$$A(z) = 1 - \sum_{i=1}^{p} a_i z^{-i}, \tag{4}$$

where a_i, $i = 1, \ldots, p$, are LP coefficients.

Linear Spectral Pairs (*LSP*) is another representation of the coefficients of the inverse filter A(z), where the p zeros of A(z) are mapped onto the unit circle in the Z–plane through a pair of auxiliary p+1-order polynomials P(z)(symmetric) and Q(z) (asymmetric)[8]:

$$P(z) = A(z) + z^{-(P+1)} A(z^{-1}) \tag{5.1}$$

$$Q(z) = A(z) - z^{-(P+1)} A(z^{-1}) \tag{5.2}$$

The *LSP*s are the frequencies of the zeros of P(z) and Q(z).

Previous researches have shown that *LSP* has explicit difference in each audio class [10]. Now, we would like to find a method to measure the *LSP* distance between two one-second audio clips.

Suppose *LSP* vector in a one-second audio clip is Gaussian, its probability distribution function (*pdf*) can be represented as following:

$$p_{LSP}(\xi) = \frac{1}{(2\pi)^5 \left| \hat{C}_{LSP} \right|^{1/2}} \exp\left\{ -\frac{1}{2} (\xi - \hat{u}_{LSP})^T \hat{C}_{LSP}^{-1} (\xi - \hat{u}_{LSP}) \right\} \tag{6}$$

where $\hat{C}_{LSP}$ is the estimated *LSP* covariance matrix and $\hat{u}_{LSP}$ is the estimated mean vector. Then the *LSP* distance between two audio clips can be defined as [8],

$$D = \int_{\xi} [p_{LSP}(\xi) - p_{SP}(\xi)] \ln \frac{p_{LSP}(\xi)}{P_{SP}(\xi)} d\xi \tag{7}$$

Under the assumption that feature *pdf*s are n-variable normal populations, (7) can be derived into,

$$\begin{aligned} D &= \frac{1}{2} tr[(\hat{C}_{LSP} - C_{SP})(C_{SP}^{-1} - \hat{C}_{LSP}^{-1})] \\ &+ \frac{1}{2} tr[(C_{SP}^{-1} + \hat{C}_{LSP}^{-1})(\hat{u}_{LSP} - u_{SP})(\hat{u}_{LSP} - u_{SP})^T] \end{aligned} \tag{8}$$

In our scheme only the covariance part of (8) is used [8]. Hence, the distance measure is defined by

$$D = \frac{1}{2} tr[(\hat{C}_{LSP} - C_{SP})(C_{SP}^{-1} - \hat{C}_{LSP}^{-1})] \tag{9}$$

This dissimilarity measure is effective to discriminate speech and noisy speech from music. Figure 5 shows an example of *LSP* distance between our audio data and speech model, obtained from our training data. The speech segment is from 0 to 200 seconds, the music segment is from 201 to 350 seconds and the noisy-

speech segment is from 351 to 450 seconds. Obviously, the *LSP* distance is different among these classes. The distance between speech data and speech model is smallest; while the distance between music and speech model is the largest.

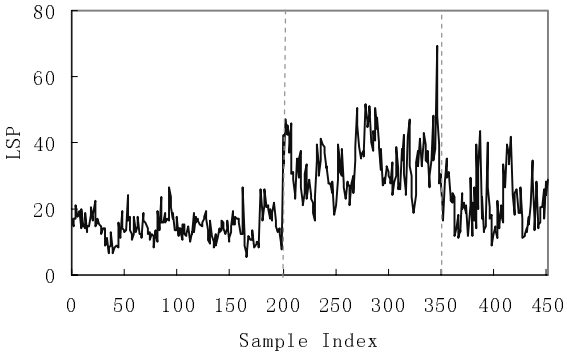

Figure 5. The LSP curve (0-200s is speech; 201-350s is music and 351-450s is noisy speech)

2.5 Band Periodicity

Band periodicity (*BP*) is defined as the periodicity of each sub-band. It can be derived by sub-band correlation analysis. Here we choose four sub-bands, they are 500~1000Hz, 1000~2000Hz, 2000~3000Hz, and 3000~ 4000Hz, respectively. The periodicity property of each sub-band can be represented by the maximum local peak of the normalized correlation function. For example, for a sine wave, its *BP* is 1; but for white noise, its *BP* is 0.

We denote the maximum local peak as $r_{i,j}(k_p)$, where k_p is the index of the maximum local peak, *i* is the band index and *j* is the frame index. It means, $r_{i,j}(k_p)$ is band periodicity of the *i*th sub-band of the *j*th frame.

To make the algorithm robust, the DC-removed full-wave regularity signal is also used for the calculation of correlation coefficient [9]. The DC-removed full-wave regularity signal is calculated as follows. First, the absolute value of the input signal is calculated. Then, it is passed through a digital filter to get DC-removed full-wave regularity signal. The transform function of the digital filter is:

$$H(z) = \frac{1 - bz^{-1}}{(1 - az^{-1})(1 + a^{*}z^{-1})} \qquad (10)$$

where variables *a* and *b* are determined experimentally, *a** is the conjunctive of *a*. The peak of normalized correlation function of the DC-removed full-wave regularity signal is denoted as $r'_{i,j}(k)$. Thus, the band periodicity is calculated as,

$$bp_i = \frac{1}{N}\sum_{j=0}^{N-1} \max(r_{i,j}(k_p), r'_{i,j}(k_p) - c) \quad i = 1,\ldots 4 \qquad (11)$$

where bp_i is the band periodicity of *i*th sub-band, *N* is the total frame number in one audio clip, *c* is used to eliminate the smoothing effect of low pass filtering [9]. In our scheme, c is 0.1.

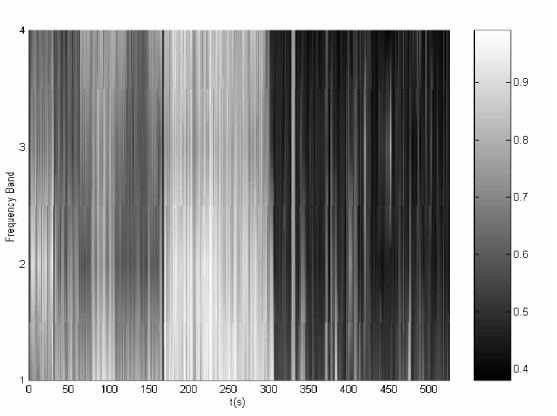

Figure 6. Band periodicity for an example audio segment. (0-150s is a music segment of tube instrument, 150-300s of piano sound, 300-520s is the concatenation of different kinds of environment sound)

Figure 6 shows an example of band periodicity comparison between music and environment sound. The music segment in the example is from 0 to 300s, while the remaining part is environment sound. It is observed that the music band periodicities are in general much higher than those of environment sound.

In our work, we use the sum of the four bands' periodicity, *bpSum*, and the periodicity of the first two bands, bp_1 and bp_2, to discriminate music and environment sound. In Figure 7, an example of the band periodicity of the music and environment sound is illustrated. The three dimensions are bp_1(500-1000Hz) and bp_2 (1000-2000Hz) and *bpSu*m. It is seen that the band periodicity of music is greater than that of environment sound in most cases, though there also exist some overlaps. This is because that music is more harmonic while environment sound is more random. Therefore, *band periodicity* is an effective feature in music and environment sound discrimination.

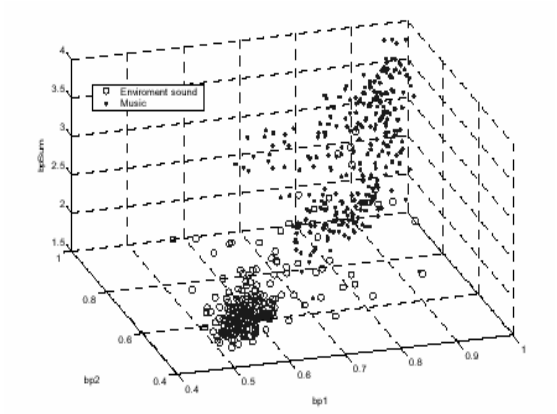

Figure 7. Band periodicity scatter graph for a piece of piano music and plane periodic noise.

2.6 Noise Frame Ratio

Noise frame ratio (NFR) is defined as the ratio of noise frames in a given audio clip. A frame is considered as a noise frame if the

maximum local peak of its normalized correlation function is lower than a pre-set threshold. The *NFR* value of noise-like environment sound is higher than that for music, because the number of noise frame of the previous class is much more, as illustrated in Figure 8.

Figure 8 shows the probability distribution curves of *NFR* for music and environment sound, which are based on our database. For music, almost no *NFR* value is above 0.3; however, for environment sound, some values are higher than 0.3, or much higher. *NFR* is really depending on how noisy the signal is. Data shows some environment sound is more noise-like.

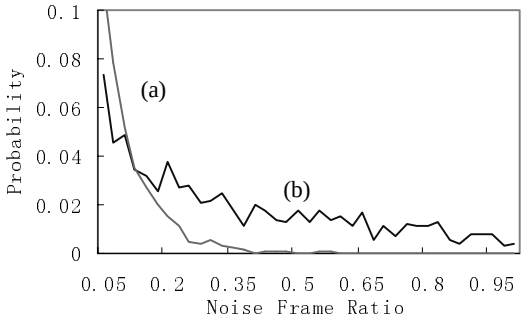

Figure 8. The probability distribution curves of *NFR*; (a) music and (b) environment sound

3. CLASSIFICATION AND SEGMENTATION SCHEME

With the features presented in the last section, a two-step scheme is proposed to classify audio clips into one of four audio classes: speech, music, environment sound and silence. First, the input audio stream is classified into speech and non-speech segments by a K-Nearest-Neighbor (KNN) classifier and Linear Spectral Pairs - Vector Quantization (LSP-VQ) analysis. Second, non-speech segments are further classified into music, environmental sound and silence, by a rule-based scheme. This two-step scheme is suitable for different application, and it can achieve a high classification accuracy.

Then, the segmentation of an audio stream can be got, by using these classification results. Some post-processing scheme is applied to further prevent misclassification.

The detail system block diagram of the proposed audio classification and segmentation scheme is shown in Figure 9.

In extracting audio features in our classification scheme, whatever the sample rate of input signal could be, we all down-sample it into 8KHz sample rate and then segment it into sub-segments by one-second window. This one-second audio clip is taken as the basic classification unit in our algorithms. It is further divided into forty 25ms non-overlapping frames, on which a 15Hz bandwidth expansion is applied. Each feature is extracted based on these forty frames in one-second audio clip. We use those features presented in the last section to represent the characteristics of each one-second audio clip.

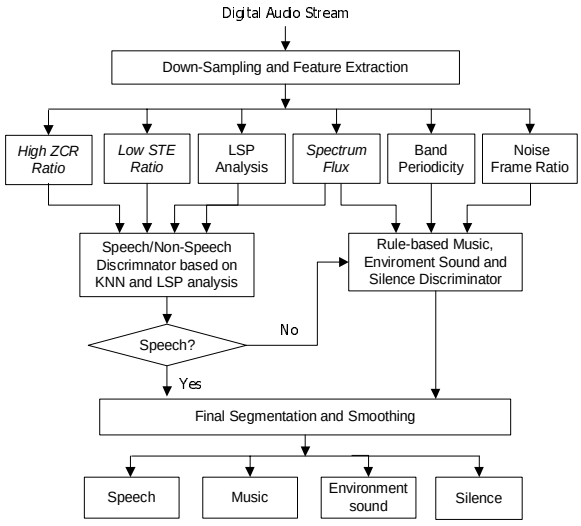

Figure9. Audio classification and segmentation system diagram

3.1 Speech/non-speech discrimination

The first step of our audio classification scheme is to discriminate speech and non-speech segments. In this scheme, we first apply a KNN classifier based on *high zero-crossing rate ratio* (*HZCRR*), *low short-time energy ratio* (*LSTER*) and *spectrum flux* (*SF*) to perform a fast pre-classification of speech and non-speech. Then, in order to refine the classification results and make the final decision, we propose a refine scheme based on *LSP distance* analysis [8]. This scheme can get higher accuracy than just combining each feature, from our experiments.

3.1.1 Pre-classification

Because of their discrimination power and low computational cost, we use *high zero-crossing rate ratio*, *low short-time energy ratio* and *spectrum flux* to form the feature vector, {*HZCRR*, *LSTER*, SF}, for fast pre-classification process.

Supposing the feature vector satisfies Gaussian mixture model, we have generated some speech codebooks and non-speech codebooks based on our training database. The training data for codebook generation is from 4 audio sequences (about two hours) of MPEG-7 test set CD1 and other 100 environment sound clips, each about 4s long. A *KNN* classifier is used in our scheme to perform audio pre-classification. In our algorithm, we use *k=2*.

It is observed that this pre-classification scheme works well in most cases and it is very fast. However, this algorithm becomes problematic when applied to signals of mixed types of audio. In fact, simple features such as *HZCRR*, *LSTER* and *SF* just characterize the fluctuations of zero-crossing rate, short-time energy and spectrum. If we add noise to speech, the fluctuation of these features for this kind of speech will become closer to that of music. Further more these features of some music with the drum background as well as some environment sounds are often similar to those of speech. Therefore, pre-classifier alone can not

assure high classification accuracy in the mixed signals. To solve this problem, we proposed a refining scheme to improve per-classification result.

3.1.2 Refining scheme

As presented in Section 2, *LSP* is a robust feature in the noisy environment for effective discrimination between noisy speech and from music. Therefore, we use this feature in refining the pre-classification results. In our scheme, we obtain a speech *LSP* covariance matrix model by training, and then save it as a speech codebook. Then we compare the distance between the speech codebook and the *LSP* covariance matrix of the testing audio clip. If the distance is smaller than a threshold, it is estimated as speech, otherwise, it is non-speech. The distance measure is defined by (9).

The procedures for final classification of speech and non-speech are illustrated in Figure 10.

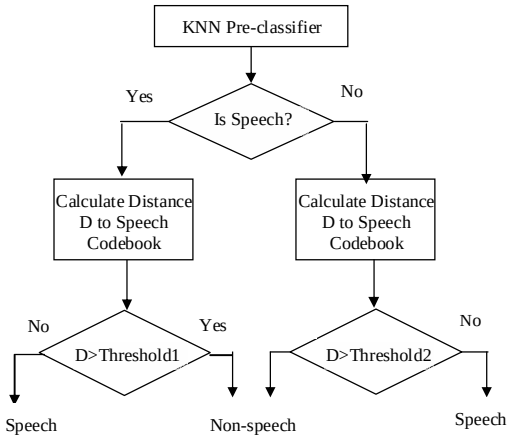

Figure 10. Final speech/non-speech discrimination

As shown in Figure 10, the result of pre-classification is examined by measuring the distance of the audio clip from the speech model codebook. We denoted the distance as D. Depending on the pre-classification result, two thresholds are used in making the final decision. If the pre-classification result is speech, then *D* is compared against *Threshold*1; then, if *D* is greater than *Threshhold*1, the audio clip is classified as non-speech. Otherwise, if the pre-classification result is non-speech, *D* is compared against *Threshhold2*, and the same rule is used to make final decision. In general, *Threshold*1 is greater than *Threshold*2. Hence, we can prevent too many pre-classification results from being converted.

In practical applications, a single speech model seems insufficient. In our scheme, we generated four speech model codebooks from training data using the LBG (Linde-Buzo-Gray) algorithm [11]. The training data include speech by different speakers of different ages and genders, in different conditions (e.g., different environment noises), etc. The dissimilarity of a test audio clip is defined as the minimum distance between the clip and the four speech model codebooks.

3.2 Music, Environment Sound, and Silence Classification Scheme

Non-speech class is further classified into music, environment sound and silence segments. In our scheme, silence detection is performed first. Then, for non-silence segment, it is classified into music or environment sound, using another set of rules.

3.2.1 Detecting silence

Silence detection is performed first based on short-time energy and zero-crossing rate in one-second window. If the average short-time energy and zero-crossing rate are lower than a threshold, the segment is classified as silence; otherwise, it is classified as non-silence segment. This simple scheme works well in our experiments.

3.2.2 Discriminating music from environment sound

Band periodicity(BP), *spectrum flux (SF)* and *noise frame ratio (NFR)* are used to discriminate music from environment. *BP* acts as the basic measure. As shown in Figure 7, the band periodicity of music is greater than that of environment sound in most cases. However, it is noted that there are some overlaps in the distribution of this feature between music and environment sound, which may lead to potential errors in the classification. To solve this problem, *SF* and *NFR* are also used.

From Figure 4, the *SF* of environment sound is much higher than that of music in many cases; and from Figure 8, almost no *NFR* value of music is higher than 0.35. Hence, these facts are utilized well in our algorithm according to the following rules.

First, if any of the bp_1, bp_2 or *bpSum* of an audio clip is lower than predefined thresholds, the clip is considered as environment sound. Otherwise, it goes to next step. The thresholds could be properly low to prevent music from being classified as environment sound.

Then, if *NFR* of a clip is greater than a given threshold, the clip is classified as noise-like environment sound. Otherwise it goes to next step again.

Third, *SF* is examined. If the *SF* is greater than a threshold, a clip is also classified as environment sound. This rule is especially useful for some strong periodicity environment sounds such as tone signal, whose *BP* and *NFR* are similar to music. Only *spectrum flux* can distinguish them.

Last, for music, its *BP* is higher, but *NFR* and *SF* are lower. It can be segmented out just by excluding all above conditions.

In our scheme, all thresholds in the rule are based on experiments.

3.3 Final Segmentation and Smoothing

Final segmentation of an audio stream can be achieved by classifying each one-second window into different audio class. Meanwhile, considering that the audio stream is always continuous in video program, it is almost impossible to change the audio types suddenly and frequently. Under this assumption, we apply some smoothing rules in final segmentation of an audio sequence. For example, if we detect a pattern of consecutive one-second windows like “speech-music-speech”, it is most likely the sequence should be all speeches, which will hence be

segmented all as speech. This smoothing process can also further prevent some misclassification.

4. EXPERIMENT RESULTS

The evaluation of the proposed audio classification and segmentation algorithms have been performed by using our database, which is gathered from MPEG-7 data set CD1 and some news and movie clips as well as some audio clips from the Internet. This database includes speech in different conditions, such as in TV studio, speeches with telephone (4kHz) bandwidth and 8kHz bandwidth. The music content in this data set is mainly songs and most of them are pop music. Such music contents are difficult for most audio classifiers.

All data are 32kHz sample rate, mono channel and 16bit per sample. From which, we select about 2 hours data as training data, and 4 hours as testing data. More in detail, the testing data includes 9587 seconds speech, 3417 seconds music and 1201 seconds environment sound. In the speech clip, the ratio between the number of pure speech and noisy speech is about 9:1. In our experiments, we set one second as a test unit.

We first implemented a baseline system which only uses the feature (*HZCRR*, *LSTER*, *SF*) with Clustering and *KNN* method, just as the 3.1.1 mentioned. The performance is shown in the following table:

Table 1. Speech, music, environment sound classification result on baseline system (unit: 100%)

Sound Type	Total Number	Discrimination Results		
		Speech	Music	ENV Sound
Speech	100	95.46	2.81	1.73
Music	100	5.24	88.39	6.37
Environment Sound	100	15.25	22.87	61.88

This baseline system works well for speech/non-speech discrimination but it doesn't work well on environment sound discrimination. So we just use it as a pre-classification method for speech and non-speech classification. More improvements are needed for environment sound classification. In our experiments, we also found the baseline system has a better performance on pure speech than noisy speech, just as the Table 2 shows.

Table 2. Baseline classification result on pure speech and noisy speech (unit: 100%)

Sound Type	Total Number	Discrimination Results	
		Speech	Music
Pure Speech	100	96.74	3.26
Noisy speech	100	73.62	26.38

In the Table 2, it could be seen that 3.26% of pure speech is detected as music, while 26.38% noisy speech is classified as music. This is because some features of noisy speech are very like those of music. So, new feature are used to increase the classification performance of noisy speech. After the refinement scheme using *LSP* distance, the performance is improved significantly, just as the Table 3 shows.

Table 3. Classification result on pure speech and noisy speech after Refinement (unit: 100%)

Sound Type	Total Number	Discrimination Results	
		Speech	Music
Pure Speech	100	98.23	1.77
Noisy speech	100	85.18	14.82

It could be seen that the accuracy for noisy speech discrimination is increased from 73.62% to 85.18%. After using our music and environment classification scheme, the accuracy for environment classification is improved from 61.8% to 79.27%. The total performance of our system is showed in Table 4.

Table 4. Speech, music, environment sound classification result before smoothing (unit: 100%)

Sound Type	Total Number	Discrimination Results		
		Speech	Music	ENV Sound
Speech	100	96.73	1.89	1.38
Music	100	3.68	91.34	4.98
Environment Sound	100	11.49	9.24	79.27

Considering the continuity of audio stream, a smoothing scheme is processed. The performance has been further improved, which is shown in Table 5.

Table 5. Speech, music, environment sound classification result (unit: 100%)

Sound Type	Total Number	Discrimination Results		
		Speech	Music	ENV Sound
Speech	100	97.45	1.55	1.00
Music	100	3.16	93.04	3.80
Environment Sound	100	10.49	5.08	84.43

From Table 5, we can see that speech, music and environment sound can be well discriminated. 97.45% Speech can be discriminated correctly; only 1.55% speech is classified into music while 1.00% is into environment sound, falsely. The total accuracy of discriminating these three classes is as high as 96.51%. If only speech and music are considered, the accuracy reaches 98.03%. The accuracy results of different discrimination type are listed in Table 6.

Table 6. The accuracy result for different discrimination type

Discrimination Type	Accuracy
Speech/music	98.03%
Speech/music/environment sound	96.51%

The experiment has shown that the proposed scheme achieves satisfactory classification accuracy. But there are still misclassifications. The discrimination of environment sounds

from music and speech is especially difficult. We did another experiment to test our algorithm. This experiment is based on a database of 457 environment sound effect clips, which includes many kinds of sound, such as automobile, beep, air plane, city life, combat, office and house. Each clip is 2-8 seconds long. In this experiment, we use a whole clip as a discriminating unit, instead of a one-second window.

The experiment results have shown that, 16 out of 457 were falsely classified into speeches, 23 into music. Thus, its accuracy can reach up to 91.47%. The detailed results are listed in the Table 7.

It was noted that some crowd sounds such as the shouting or cheering were misclassified as music, because the human voice makes the periodicity of the sound very similar to songs. Some animal sounds were also misclassified into speech, because of the similarity between the vocal track shape of some animals and that of human being. However, in general, the classification accuracy is satisfactory.

Table 7. Environment sound discrimination accuracy

Sound Type	Testing Sample Number	Error Discrimination Number
Aviation	8	0
Animals	45	2(M), 4(S)
Autos	17	1(M)
Beep	54	3(M)
Cartoon	66	1(M), 6(S)
Casino	11	2(M)
City life	115	6(M), 2(S)
Combat	21	0
Crowds	14	3(M)
House	24	1(S)
Office	32	0
Others	50	5(M),3(S)
Totals	457	23(M),16(S)

We have also tested the time complexity of our algorithm. With Pentium III 667MHz PC/Windows 2000, the segmentation and classification process can be completed in about 15% of the time-length of a audio clip. The correlation calculation in computing LSP matrix and band periodicity is the most time-consuming part in our algorithm. After using an optimized function to compute these two features, the time performance has been increased dramatically. Our scheme can totally suit the real-time processing in multimedia application.

5. CONCLUSIONS

In this paper we have presented our study on audio content analysis in the context of video browsing. We have described in detail a novel two-stage audio segmentation and classification scheme that segments and classifies an audio stream into speech, music, environment sound and silence. These classes are the basic data set for video structure extraction. A novel two-stage algorithm has been developed and presented. The first stage of the classification is to separate speech from non-speech, based on simple features such as high zero-crossing rate ratio, low short-time energy ratio, spectrum flux and *LSP* distance. The second stage of the classification further segments non-speech class into music, environment sounds and silence with a rule based classification scheme. In this process, we introduced two new features: *noise frame ratio* and *band periodicity*. Experimental evaluation has shown that the proposed audio classification scheme is very effective and the total accuracy rate is over 96%.

In the future, we will improve our classification scheme to discriminate more audio classes. We will also focus on developing an effective scheme to apply audio content analysis to improve video structure parsing and indexing process.

6. REFERENCES

[1] J. Foote. *Content-based retrieval of music and audio*. In C. C. J. Kuo et al., editors, Multimedia Storage and Archiving Systems II, Proc. of SPIE, volume 3229, pages 138-147, 1997.

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

[3] S. Pfeiffer, S. Fischer and W. Effelsberg. *Automatic Audio Content Analysis*, Proceedings of the fourth ACM international conference on Multimedia, pp. 21-30, 1996.

[4] J. Saunders. *Real-time Discrimination of Broadcast Speech/ Musi*c. Proc. ICASSP96, vol.II, pp.993-996, Atlanta, May, 1996

[5] E. Scheirer and M. Slaney, *Construction and Evaluation of a Robust Multifeature Music/Speech Discriminato*r. Proc. ICASSP 97, vol. II, pp 1331-1334. IEEE, April 1997

[6] D. Kimber and L. Wilcox. *Acoustic Segmentation for Audio Browsers*, Proc. Interface Conference, Sydney, Australia, July, 1996

[7] T. Zhang and C.-C. J. Kuo. *Video Content Parsing Based on Combined Audio and Visual Information*. SPIE 1999, Vol. IV, pp. 78-89, 1999.

[8] J. P. Campbell, JR. *Speaker Recognition: A Tutoria*l. Proceedings of the IEEE, vl.85, no.9, pp.1437~1462, 1997.

[9] A. V. McCree and T. P. Barnwell. *Mixed Excitation LPC Vocoder Model for Low Bit Rate Speech Coding*. IEEE Transaction on Speech and Audio Processing, vol. 3, No. 4, pp242-250. July 1995.

[10] K. El-Maleh, M. Klein, G. Petrucci and P. Kabal. *Speech/music discrimination for multimedia application*. ICASSP00, 2000

[11] Y. Linde, A. Buzo, and R.M. Gray. *A Algorithm for Vector Quantizer Desig*n, IEEE Trans. on Comm. Com-28, No.1, pp. 84-95, 1980.

[12] S. Srinivasan, D. Petkovic and D. Ponceleon. *Towards robust features for classifying audio in the CueVideo System*. Proceedings of the seventh ACM international conference on Multimedia, pp.393 – 400, 1999.

[13] Z. Liu, Y. Wang and T. Chen. *Audio Feature Extraction and Analysis for Scene Segmentation and Classification*. Journal of VLSI Signal Processing Systems, June 1998

[14] J. S. Boreczky and L. D. Wilcox. *A Hidden Markov Model Frame Work for Video Segmentation Using Audio and Image Features*. Proceedings of ICASSP'98, pp.3741-3744, Seattle, May 1998.