Robust Features for Effective Speech and Music Discrimination

Zhong-hua Fu[1], Jhing-Fa Wang[2]

School of Computer Science
Northwestern Polytechnical University, Xi'an, China[1]
Department of Electrical Engineering
National Cheng Kung University, Tainan, Taiwan[1, 2]
mailfzh@mail.ncku.tw[1], wangjf@csie.ncku.edu.tw[2]

Abstract

Speech and music discrimination is one of the most important issues for multimedia information retrieval and efficient coding. While many features have been proposed, seldom of which show robustness under noisy condition, especially in telecommunication applications. In this paper two novel features based on real cepstrum are presented to represent essential differences between music and speech: Average Pitch Density (APD), Relative Tonal Power Density (RTPD). Separate histograms are used to prove the robustness of the novel features. Results of discrimination experiments show that these features are more robust than the commonly used features. The evaluation database consists of a reference collection and a set of telephone speech and music recorded in real world.

Keywords: Speech/Music Discrimination, Multimedia Information Retrieval, Real Cepstrum.

1. Introduction

In applications of multimedia information retrieval and effective coding for telecommunication, audio stream always needs to be diarized or labeled as speech, music or noise or silence, so that different segments can be implemented in different ways. However, speech signals often consist of many kinds of noise, and the styles of music such as personalized ring-back tone may differ in thousands ways. Those make the discrimination problem more difficult.

A variety of systems for audio segmentation or classification have been proposed in the past and many features such as Root Mean Square (RMS) [1], Zero Crossing Rate (ZCR) [1,4,5], low frequency modulation [2,4,5], entropy and dynamism features [2,3,6], Mel Frequency Cepstral coefficients (MFCCs) have been used. Some features need high quality audio signal or refined spectrum detail, and some cause long delay so as not fit for telecommunication applications. While the classification frameworks including nearest neighbor, neural network, Hidden Markov Model (HMM), Gaussian Mixture Modal (GMM) and Support Vector Machine (SVM) have been adopted as the back end, features are still the crucial factor to the final performance. As shown in the following part of this paper, the discrimination abilities of some common features are poor with noisy speech. The main reason may explain as that they do not represent the essential difference between speech and music.

In this paper, two novel features, called as Average Pitch Density (APD) and Relative Tonal

Power Density (RTPD) are proposed, which are based on real cepstrum analysis and show better robustness than the others. The evaluation database consists of two different data sets: one comes from Scheirer and Slaney [5], the other is collected from real telecommunication situation. The total lengths for music and speech are about 37 minutes and 28.7 minutes respectively.

The rest of this paper is organized as follows: Section 2 introduces the novel features based on real cepstrum analysis. Section 3 describes the evaluation database and the comparative histograms of different features. The discrimination experiments and their results are given in section 4. Section 5 concludes this paper.

2. Features Based on Real Cepstrum

There are tremendous types of music, and the signal components of which can be divided into two classes: tonal-like and noise-like. The tonal-like class consists of tones played by all kinds of musical instruments, and these tones are catenated to construct the melody of music. The noise-like class is mainly played by percussion instruments such as drum, cymbal, gong, maracas, etc. The former class corresponds to the musical system, which construct by a set of predefined pitches according to phonology. The latter class can not play notes with certain pitch and is often used to construct rhythm.

The biggest difference between speech and music lies on the pitch. Because of the restriction of musical system, the pitch of music usually can only jump between discrete frequencies, except for vibratos or glissandi. But pitch of speech can change continuously and will not keep on a fixed frequency for a long time. Besides the difference of pitch character, the noise part of music, which is often played by percussion instrument, also has different features from speech. That part of music does not have pitch, but it usually has stronger power. This phenomenon seldom exists in speech signal, because generally the stronger part of speech is voiced signal, which does have pitch.

In order to describe the differences of pitch between speech and music, we use real cepstrum instead of spectrogram. Cepstrum analysis is a more powerful tool to analysis the detail of spectrum, which can separate pitch information from spectral envelop. The real cepstrum is defined as (Eq. (2) gives the Matlab expression)

$$RC_x \hat{=} real\left(\frac{1}{2}\int_{-\pi}^{\pi}\log\left|X\left(e^{j\omega}\right)\right|e^{j\omega n}d\omega\right) \tag{1}$$

$$RC_x = real(ifft(\log(abs(fft(x))))) \tag{2}$$

Where x is a frame of audio signal weighted by hamming window, of which the discrete Fourier transform is $X\left(e^{j\omega}\right)$. $real(\cdot)$ denotes extracting the real part of the complex results. RC_x are the coefficients of real cepstrum. The coefficients that near zero origin reflect the big scale information of power spectrum such as the spectrum envelop, and those far from the zero origin show the spectrum detail. Figure 1 uses the latter to demonstrate the differences of pitch between speech and music. It is clear that the music pitches are jumped discretely while speech pitches do not. Figure 2 uses spectrogram to show the noise-like feature of a rock music segment, where most ictus have no pitch.

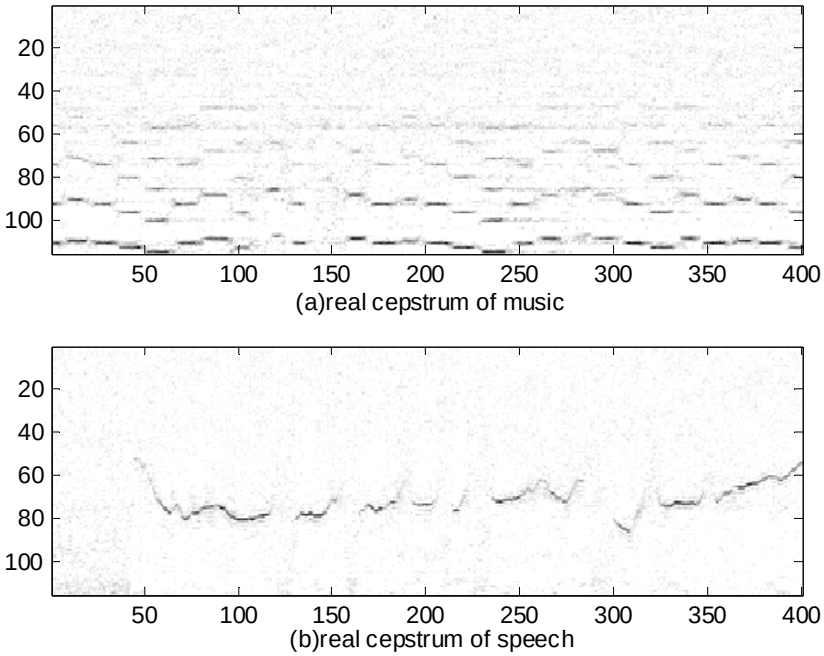


Figure 1. Pitch different between music (a) and speech (b) by means of real cepstrum. Only coefficients far from the zero origin are used.

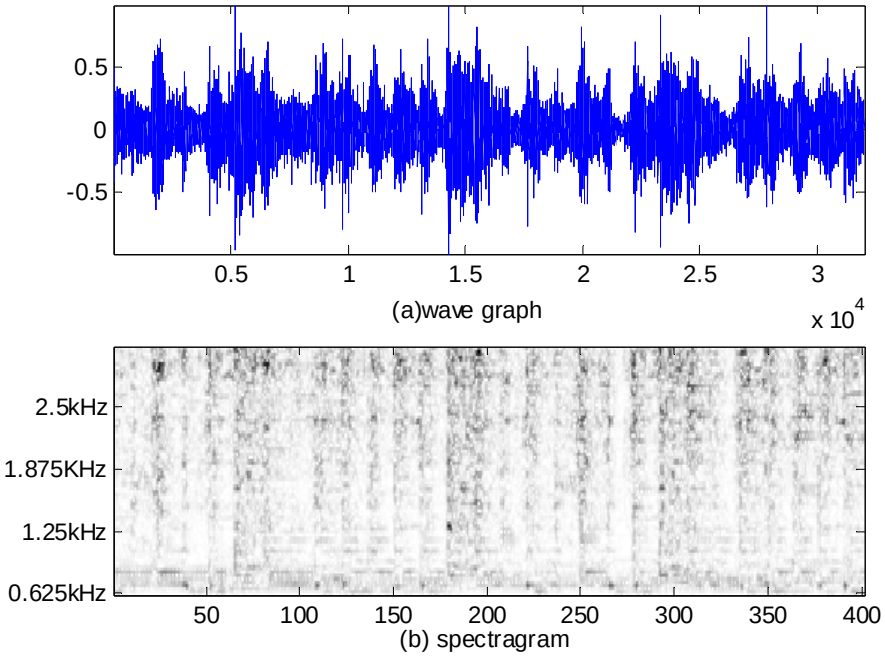


Figure 2. Waveform and spectrogram of a segment of rock music. It is clear to find that most ictus have no pitch.

To parameterize the above conclusion, we propose two novel features: Average Pitch Density (APD) and Relative Tonal Power Density (RTPD).

A. APD feature

Because of the musical instruments and polyphony, the average pitch usually is higher than speech. The APD feature is independent with signal power and reflects the details about spectrum, which is defined as

$$APD(K) = \sum_{i=K*N+1}^{K*N+N} \frac{1}{L} \sum_{j=l_1}^{l_2} \left| RCx_i(j) \right| , \text{ where } \quad L = l_2 - l_1 + 1 \tag{3}$$

where K means the K-th analysis segment, and N is the length of it. L is number of RC_x coefficients that far from zero origin, whose range is l_1 to l_2. This feature is relative simple, but it does prove to be robust for discrimination between speech and music. The histogram in figure 3 (e) demonstrate this conclusion.

B. RTPD feature

While the detail information about spectrum can be used to discriminate tonal or song from speech, the variation of energy combined with pitch information may be used to separate percussive music from noisy speech. In clean or noisy speech signal, the segments that show clear pitch usually are voiced speech, which are likely to have bigger energy. So if all segments with pitch are labeled as tonal parts and the others are label as non-tonal parts, we can probably say that if the energy of tonal parts is smaller than that of non-tonal parts, then the segment may not be speech, otherwise the segment can be speech or music.

In order to label tonal and non-tonal parts, we still use real cepstrum. Since if clear pitch does exist, a distinct stripe will appear in real cepstrum, even if in noise condition. We use the peak value of *RCx* that far from zero origin to judge tonal or non-tonal. The threshold we choose is 0.2. Frames whose peak value is bigger than 0.2 are labeled as tonal, or else are labeled as non-tonal. Thus the RTPD can be defined as

$$RTPD(K) = \underset{i \in \Theta}{mean}\left(RMS_i\right) \Big/ \underset{j \in \Psi}{mean}(RMS_j) \tag{4}$$

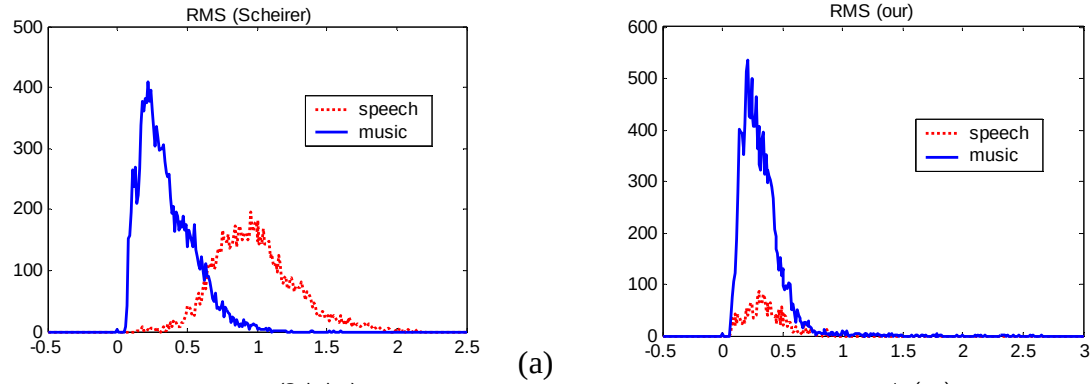
where Θ consists of all tonal frames of K-th analysis segment, and Ψ is the entire set of frames of the segment. RMS_i is the root mean square of the i-th frame.

3. Discrimination Ability

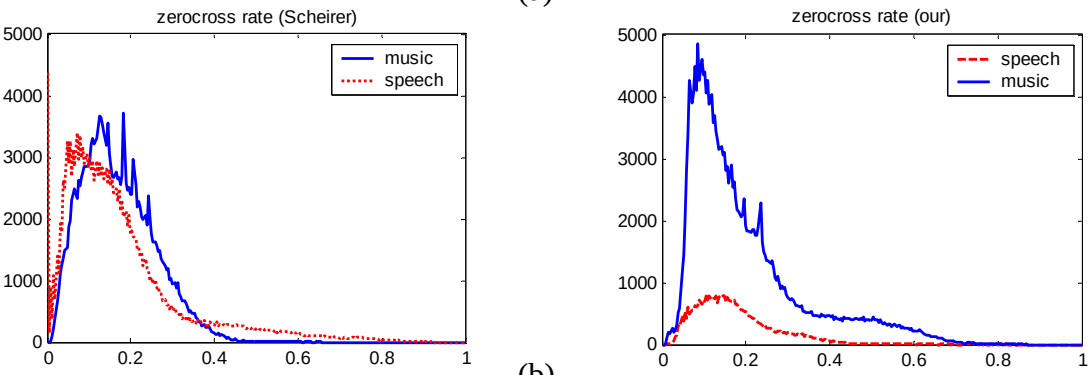
Due to the lack of a standard database for evaluation, the comparisons between different features are not easily. Our evaluation database consists of two parts: one comes from collection of Scheirer and Slaney[5], the other comes from the real records from telecommunication application. The former includes speech sets and music sets. Each set contains 80 15-second long audio samples. The samples were collected by digitally sampling an FM tuner (16-bit monophonic samples at a 22.05 kHz sampling rate), using a variety of stations, content styles, and noise levels. They made a strong attempt to collect as much of the breadth of available input signals as possible (See [5] for details). The latter set is recorded by us based on telecommunication application, which has 25 music files and 174 noisy speech files, 17 and 11.7 minutes in length respectively. Especially, the speech signals of the latter set consist of many kinds of live noises, which are non-stationary with different SNR.

Based on the two data sets above, we build an evaluation corpus by concatenating those files

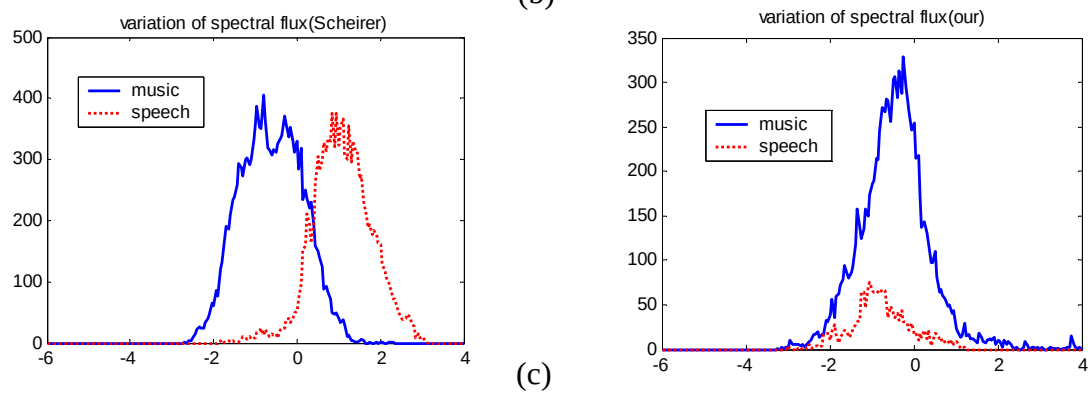
randomly into two columns: CLN-Mix and ZX-Mix. CLN-Mix contains 20 mixed files, each concatenates 2 speech samples and 2 music samples which are all extracted from Scheirer's database. ZX-Mix uses the same way except that all samples are chosen from our records. With these databases, we compared 4 commonly used features with our prompted ones. They are (1) RMS; (2)zero crossing rate; (3)variation of spectral flux; (4)percentage of "low-energy" frames. Figure 3 shows the discrimination abilities of each feature with Scheirer's and our database. It is clear that those 4 features show poor performance in noise situation, while APD and RTPD show more robust



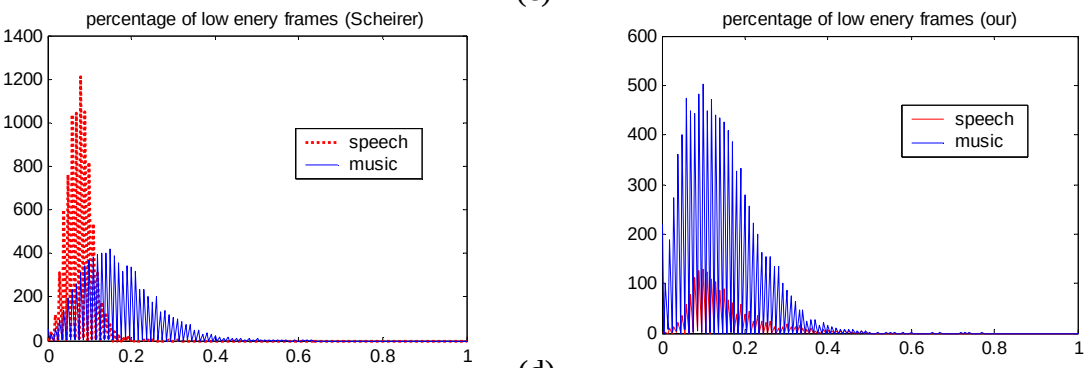
(a)



(b)



(c)



(d)

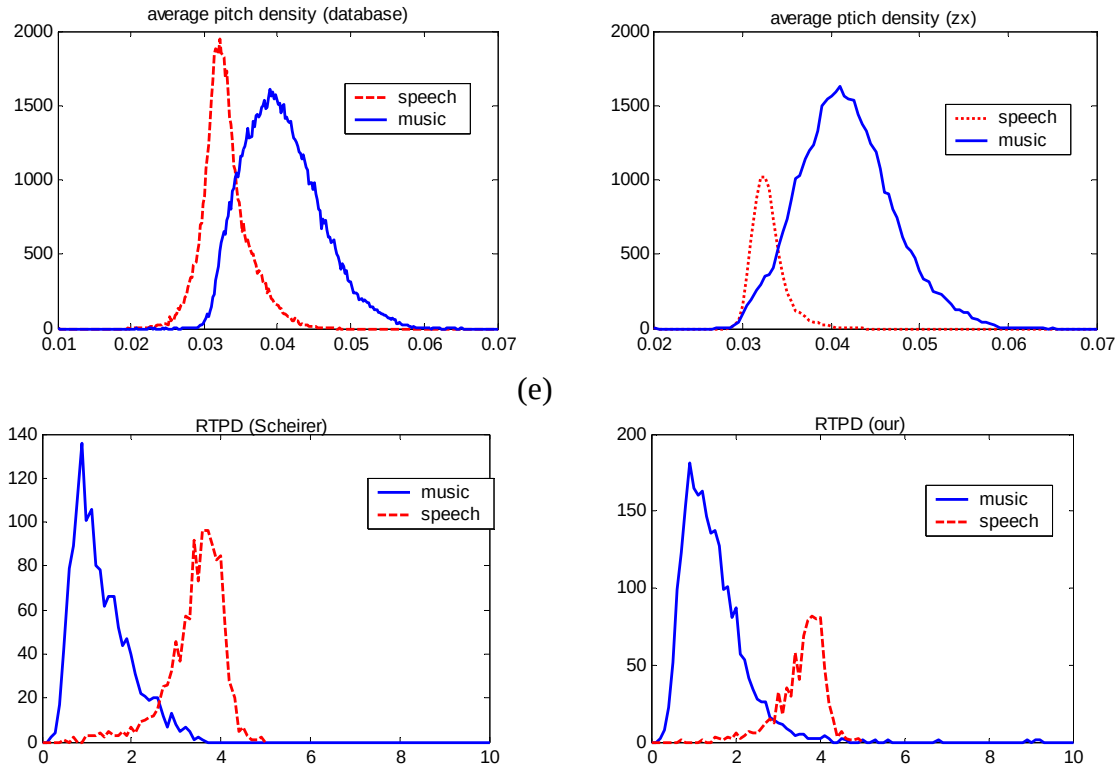


Figure 3. Histograms of different features for speech/music discrimination. (a)-(f) are RMS, ZCR, variation of spectral flux, percentage of "low-energy" frames, APD, RTPD.

4. Discrimination Experiments

In many speech and music discrimination system, GMM is commonly used for classification. A GMM models each class of data as the union of several Gaussian clusters in the feature space. This clustering can be iteratively derived with the well-known EM algorithm. Usually the individual clusters are not represented with full covariance matrices, but only the diagonal approximation. GMM uses a likelihood estimate for each model, which measurers how well the new data point is modeled by the entrained Gaussian clusters.

We use 64 components GMM to modal speech and music signal separately. The feature vector consists of: (1) APD; (2) RTPD; (3) log of variance of RMS; (4) log of variance of spectral centroid; (5) log of variance of spectral flux; (6) 4Hz modulation energy; (7) dynamic range. Training data consists of the training part of Scheirer's database and 8 minutes of noisy speech recorded. CLN-Mix and ZX-Mix are used for evaluation.

The frame length is 10ms, and the analysis windows for proposed features extraction is 1 second (100 frames) with 10 new input frames each time. For comparison, MFCC + delta + acceleration (MFCC_D_A) feature for each frame is also examined. GMM with 64 mixtures is used for speech and music respectively. For classification, every proposed feature vector is used to calculate the log likelihood score, and correspondingly, 10 frames MFCC_D_A features are used. The experimental results are list in Table 1. Furthermore, we also use the adjacent 10 proposed feature vectors for one decision and 100 frames of MFCC_D_A features are used as well. The results are shown in Table 2.

It is clear that MFCC _D_A features have good ability for discrimination with CLN-Mix data, but drop distinctly with ZX-mix, especially for music signals. But on both data sets, our

proposed features work well and express robustness in noise condition.

Table 1. Speech/Music Discrimination Accuracies in Every 100ms

Accuracy	MFCC_D_A		Proposed	
	Speech	Music	Speech	Music
CLN-Mix	91.56%	89.81%	93.78%	91.48%
ZX-Mix	99.91%	64.41%	94.19%	93.13%

Table 2. Speech/Music Discrimination Accuracies in Every Second

Accuracy	MFCC_D_A		Proposed	
	Speech	Music	Speech	Music
CLN-Mix	93.98%	95.11%	95%	92.86%
ZX-Mix	100%	67.39%	100%	94.45%

5. Conclusion

Two novel features have been presented in this paper for robust discrimination between speech and music, named Average Pitch Density (APD) and Relative Tonal Power Density (RTPD). As shown in separate histograms, many other commonly used features do not work in noisy condition, but the novels show more robustness. When combined with the other 5 robust features, the accuracies of discrimination are higher than 90%. The results mean that the novel features may represent some essential differences between speech and music.

There are many interesting directions in which to continue pursuing this work. Since the real cepstrum can show many differences between speech and music, there will be other novel features which represent the holding and changing characters of pitches. What's more, more researches are needed for better classification and feature combinations.

References

[1] C. Panagiotakis, G. Tziritas, *A Speech/Music Discriminator Based on RMS and Zero-Crossings*, IEEE Transactions on Multimedia, Vol.7(1), February 2005.

[2] O. M. Mubarak, E. A. Ambikairajah, J. Epps, *Novel Features for Effective Speech and Music Discrimination*, Proc. IEEE International Conference on Engineering of Intelligent Systems, pp.1-5, April 2006.

[3] J. E. Muñoz-Expósito , S. García-Galán , N. Ruiz-Reyes , P. Vera-Candeas, *Adaptive Network-based Fuzzy Inference System vs. Other Classification Algorithms for Warped LPC-based Speech/Music Discrimination*, Engineering Applications of Artificial Intelligence, Vol. 20(6), pp.783-793, September, 2007.

[4] M. J. Carey, E. S. Parris, H. Lloyd-Thomas, *A Comparison of Features for Speech, Music Discrimination*, Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol.1, pp. 149-152, March 1999.

[5] E. Scheirer, M. Slaney, *Construction and Evaluation of a Robust Multifeature Speech /Music Discriminator*, Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol.1, pp. 1331-1334, April 1997.

[6] T. Zhang, J. Kuo, *Audio Content Analysis for On-line Audiovisual Data Segmentation and Classification*, IEEE Transactions on Speech Audio Processing, Vol. 9 (3), pp. 441-457, May 2001.