FEATURE EXTRACTION FOR SPEECH AND MUSIC DISCRIMINATION

Huiyu Zhou, Abdul Sadka and Richard M. Jiang

Brunel University, Uxbridge, Middlesex, United Kingdom
E-mail:{Huiyu.Zhou, Abdul.Sadka, Min.Jiang@brunel.ac.uk}

ABSTRACT

Driven by the demand of information retrieval, video editing and human-computer interface, in this paper we propose a novel spectral feature for music and speech discrimination. This scheme attempts to simulate a biological model using the averaged cepstrum, where human perception tends to pick up the areas of large cepstral changes. The cepstrum data that is away from the mean value will be exponentially reduced in magnitude. We conduct experiments of music/speech discrimination by comparing the performance of the proposed feature with that of previously proposed features in classification. The dynamic time warping based classification verifies that the proposed feature has the best quality of music/speech classification in the test database.

1. INTRODUCTION

Video scene analysis and classification are highly demanding in information retrieval, video editing and human-computer interface. Rich literature in this field is addressed on the use of audio and/or visual components. In recent years, as one of the common research interests, the integration of audio and visual observations attracted enormous attention, where the accompanying audio information helped identify individual scenes. For example, the audio of musical events is significantly different from that of the news report, and this audio difference can be used to discriminate between these two different scenarios. To achieve this purpose, proper classification of music and speech is a necessary element in the analysis. In this paper, our major concern is the determination of appropriate features that can differentiate audio clips associated with various scene classes. Although this algorithmic development is directly linked to the classification of music and speech, it can be easily adapted to accommodate other applications, e.g. differentiation of sports events and news reporting scenarios.

Audio features can be used to characterize the media signals for discrimination between music and speech classes. In general, these audio features can be catergorised into two groups: Time and frequency domains. The former includes zero-crossing rates, amplitudes and pitches, while the latter consists of spectragrams, cepstral coefficients and Mel-frequency cepstral coefficients (MFCC), etc. Evidence shows that spectral features have demonstrated superiority to temporal ones in some of the applications. For example, Scheirer *et al.* [1] and Saad *et al.* [2] examined the following five features to measure conceptually distinct properties of speech and music signals: Percentage of low energy frames, rolloff point of the spectrum, spectral flux, zero-crossing rate, spectral centroid. The effectiveness of the spectral features has been validated, although it lacks a discussion about the importance of individual features. Carey *et al.* [3] used the following features for classification practice: Cepstral coefficients, Delta cepstral coefficients, amplitude, Delta amplitude, pitch, Delta pitch, zero-crossing rate, and Delta zero-crossing rate. They discovered that the best classification was achieved using the cepstral and Delta cepstra. El-Maleh *et al.* [4] conducted frame-level speech/music discrimination using features such as line spectral frequencies (LSF), differential LSF and successive differences of LSF, LSF with the zero crossing count, and LSF with Linear prediction zero crossing ratio plus the ratio of the zero crossing count (ZCC) of the input and the ZCC of the output. It was discovered that the K-nearest neighbor classifier with spectral features provided an optimal performance.

A large number of features in either separate or combinatorial forms have been reported in the literature for music/speech classification. Since most the established classifiers still hold significant incorrect classification rates to different test databases, we believe that there is a room for improvement of classification performance. In this work, we focus the attention on the specific problem of audio classification in music and speech, assuming that the silence segments have already been identified using, for example, a Linear Predictive Coding method proposed in [5]. Especially, our work is directed towards introducing a new feature extraction scheme rather than improving the state-of-the-art classifiers. The novel audio feature is extracted according to the multiplication of MFCC estimates and an exponential component that depends on the outcome of the MFCC estimates. To validate its effectiveness, we apply a dynamic time warping (DTW) scheme to a music/speech database for music/speech discrimination using this new feature.

This work was supported by European Commission under Grant FP6-045189-STREP (RUSHES).

978-1-4244-2044-5/08/$25.00 ©2008 IEEE.

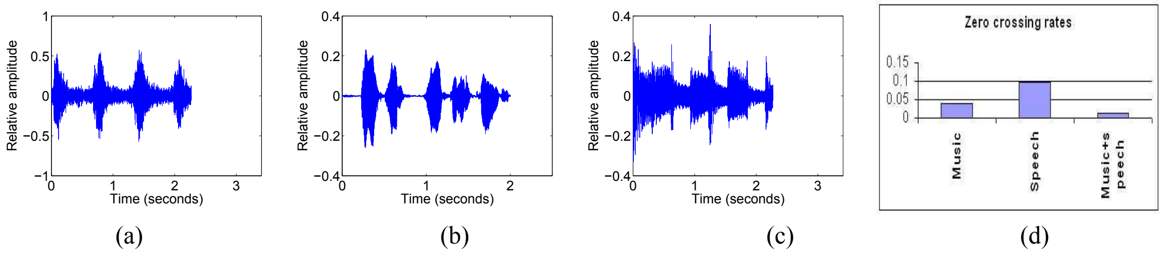


Fig. 1. Examples of audio segments: (a) music only, (b) speech only, (c) music plus speech, and (d) zero crossing rates of (a), (b) and (c).

2. AUDIO FEATURES EXTRACTION FOR MUSIC/SPEECH CLASSIFICATION

Music/speech discrimination relies on appropriate feature extraction, which can help reduce the dimensionality of unknown variable space [6]. A good feature extraction scheme can be used to present the main characteristics of individual audio classes. To enhance the classification rates, temporal features can be considered as well as spectral features [7]. Note that in this paper we intend to extract audio features in a longer term clip level (2-3 seconds).

2.1. Previously used features

To differentiate speech from music, we need to carefully look into the characteristics of these two classes. Speech appears with a regular structure where music does not show. For example, speech is composed of a succession of vowels and consonants: While the vowels are high energy events with most of the spectral energy contained at low frequencies, the consonant are noise-wise, with the spectral energy distributed more towards the higher frequency bands. Driven by these facts, most previously used features have been established in the domain of spectral analysis for music/speech discrimination. Here, we summarise some of the representatives.

Zero-crossing rate is the measurement of the number of times that the audio signal curve passes through a zero level within a speech frame. It can be severely affected by noise. Speech signals have higher zero-crossing rates than music. Linear predictive coefficients (LPC) is a method that predicts the next sample according to a weighted sum of n previous samples, i.e.

$$\tilde{s}(t) = \sum_{i=1}^{n} w_i s(t-i), \qquad (1)$$

where w_i are the weights or prediction coefficients, $s(t-i)$ represents a sample at time instance $t-i$. w_i can be determined by minimising the mean squared error between the real sample and the prediction. The linear prediction coefficients can use the Levinson-Durbin recursion to solve the normal equations that arise from the least-squares formulation [8].

Spectral flux or Delta spectrum magnitude is the measurement of frame-to-frame spectral difference so it describes the shape change of the spectrum. In the case of music/speech discrimination, music signals are often of more regular spectral variations than speech. Percentage of low energy frames (%LEF) refers to the proportion of frames with root mean-squared (RMS) power less than 50% of the mean RMS power in a given period. Mel-frequency cepstral coefficients (MFCC) is a parameter used in the discrimination due to the spectral difference between music and speech. A commonly used formula to approximately reflect the relation between the Mel-frequency and the physical frequency is given by

$$M(f) = 1125 \times \log_{10}\left(1 + \frac{f}{700}\right), \qquad (2)$$

where f is frequency. Perceptual linear prediction (PLP), similar to LPC analysis, is based on the short-term spectrum of speech. In contrast to pure linear predictive analysis of speech, perceptual linear prediction modifies the short-term spectrum of the speech by several psychophysically based transformations. Relative spectra filtering (RASTA) band passes each feature coefficient [9]. Linear channel distortions appear as an additive constant in both the log spectral and the cepstral domains. The high-pass portion of the equivalent band pass filter alleviates the effect of convolutional noise introduced in the channel. The low-pass filtering helps in smoothing frame to frame spectral changes. Music signals hold much more frequency details than speech, and therefore they can be differentiated in the domain of RASTA.

Fig. 1 shows different audio clips of music only, speech only and music plus speech. It shows that zero-crossing rates of speech only are significantly larger than the other signals. Fig. 2 denotes individually extracted features of Fig. 1(a) and (b) by using some classical techniques introduced above.

2.2. Improved MFCC

Before introducing any new feature, we need to mention the concept of Delta MFCC [3]. Delta MFCC is used to catch the differenced (or delta) cepstrum between the different frames.

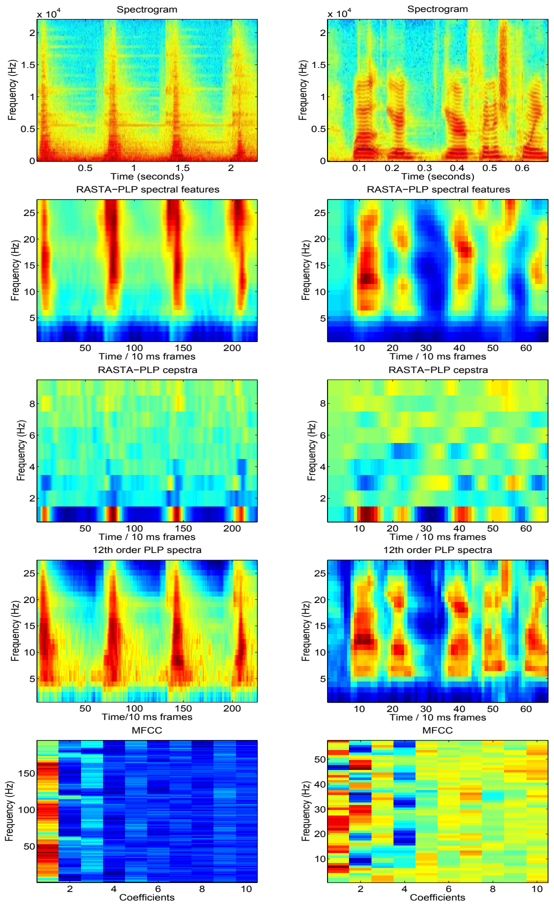


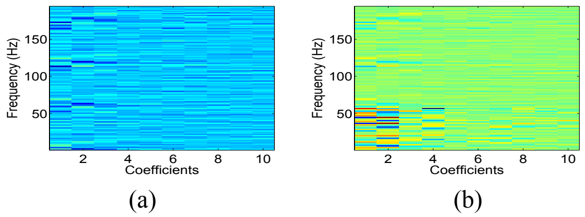
Fig. 2. Extracted audio features using classical methods: left column - music results; right column - speech results.

This variation can be defined as:

$$\Delta M(i,j) = (M(i,j+1) - M(i,j-1))/2, \tag{3}$$

where i and j denote the indices of coefficients and frequency bands of MFCC estimates M, respectively. Delta MFCC has demonstrated its optimality in music/speech classification [3]. This delta MFCC represents the linear difference between two neighboring cepstrums and hence shares the common properties with MFCC. Fig. 3 shows the delta MFCC of Fig. 1(a) and (b).

Interestingly, the areas with strong contrast in Fig. 2 attract our attention. We discover that these areas actually accompany significantly varied spectra, compared to their neighbors. That is to say, human perception tends to pay more attention to the areas whose spectra magnitudes differ from the mean of the entire spectral set. Nevertheless, this does not imply that the observer completely discards the averaging audio information (background). Without these background signals, it is impossible to identify the perceptible spectra [10].



(a) (b)

Fig. 3. Delta MFCC of music and speech: left column - music; right column - speech.

To simulate this biological procedure, we propose a new strategy for spectral analysis. In details, we attempt to extract audio features namely "improved MFCC", based on the outcomes of MFCC estimation: Firstly, we calculate the mean m of frequency bands of each coefficient. This is followed by the computation of variance σ of the frequency bands. Once this has been done, we multiply the original cepstrum data with an exponential component, resulting in a form as follows:

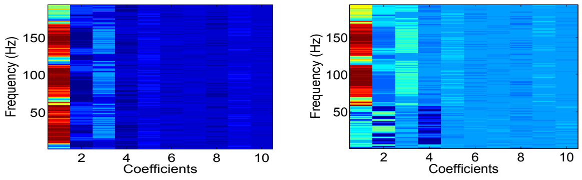
$$M_s(i,j) = M(i,j) \exp^{-\frac{(M(i,j)-m)^2}{2\sigma^2}}. \tag{4}$$

Fig. 4 illustrates the "improved MFCC" of Fig. 1(a) and (b). It is observed that Fig. 4 has a better discriminative pattern than Fig. 3 in this example. This is due to the fact that the mean cepstrum of the entire spectra is used as a threshold so the difference between two spectral values can be amplified. The closer toward the mean cepstrum, the larger MFCC magnitudes.

3. EXPERIMENTAL WORK

In this section, we evaluate the proposed "improved MFCC" feature in music/speech discrimination by comparing its performance with that of classical spectral features MFCC, Delta MFCC, RASTA-PLP cepstra, 12th order PLP spectra. Here, we apply a dynamic time warping based classifier for similarity check. Dynamic time warping [11] is used due to its capability in handling two sequences which may vary in time or speed. In general, templates of music and speech only signals are stored with their individual features computed. Test data comes to the classifier after its features are available. Then the features of test data and the templates will be checked for similarity. We use a "music-speech" corpus that is part of a collection of 240 15-second extracts collected from the radio by Scheirer [1]. The data consists of training and test data, and is further categorised as speech only, music (with or without vocals) and speech over music, which will be classified individually.

For example, Fig. 5 illustrates the performance comparison of different feature extraction methods in music/speech discrimination. The values shown in the figure indicate correct classification rates in individual cases (e.g. "music vs



(a) (b)

Fig. 4. Improved MFCC of music and speech: left column - music; right column - speech.

music+speech" means that we intend to discriminate between music and combination of music and speech). Clearly, larger values correspond to better discrimination capability as expected. We can observe that the proposed "improved MFCC" feature allows us to achieve the best discrimination quality in all these tests. In the meantime, it also exhibits that other methods cannot hold consistent performance throughout the overall tests.

4. CONCLUSIONS AND FUTURE WORK

We proposed a new audio feature for optimal music and speech discrimination, while a dynamic time warping classifier using this feature was evaluated in a number of experiments. Generally speaking, we multiplied the MFCC estimates with an exponential component in order to simulate the human perception of paying more attention to the areas of larger cepstral variations. We conducted experiments of music/speech discrimination by comparing the performance of the proposed feature with that of previously proposed features. The dynamic time warping based classification verified that the proposed feature had the best quality of music/speech classification. The future work will be addressed on the applications of the proposed audio feature in audiovisual retrieval, while it is worthy to try different classifiers with the same features on the music/speech classification.

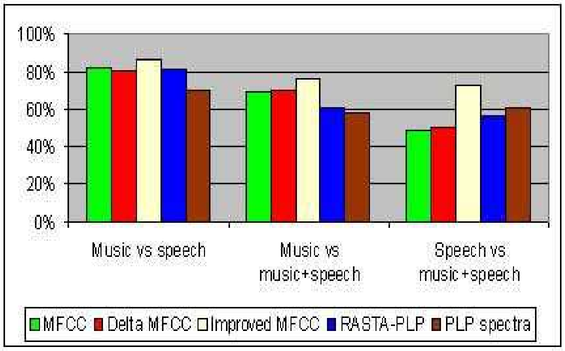


Fig. 5. Performance comparison of different feature extraction methods in music/speech discrimination, where y-axis refers to correct classification rates.

5. REFERENCES

[1] E. Scherer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proc. of the International Conference on Acoustics, Speech, and Signal Processing*, 1997, pp. 1331–1334.

[2] E.M. Saad, M.I. El-Adawy, M.E. Abu-El-Wafa, and A.A. Wahba, "A multifeature speech/music discrimination system," in *Proc. of the 19th National Radio Science Conference*, 2002, pp. 208–213.

[3] M.J. Carey, E.S. Parris, and H. Lloyd-Thomas, "A comparison of features for speech, music discrimination," in *Proc. of the International Conference on Acoustics, Speech, and Signal Processing*, 1999, pp. 149–152.

[4] K. El-Maleh, M. Klein, G. Petruci, and P. Kabal, "Speech/music discrimination for multimedia applications," in *Proc. of the International Conference on Acoustics, Speech, and Signal Processing*, 2000, pp. 2445–2448.

[5] L. Rabiner and M. Sambur, "Application of an lpc distance measure to the voiced-unvoiced-silence detection problem," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 25, no. 4, pp. 338–343, 1977.

[6] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.

[7] M.K.S. Khan and W.G. Al-Khatib, "Machine-learning based classification of speech and music," *Multimedia Syst.*, vol. 12, no. 1, pp. 55–67, 2006.

[8] G. Tzanetakis and P. Cook, "A framework for audio analysis based on classification and temporal segmentation," in *Proc. of EUROMICRO Workshop on Music Technology and Audio Processing, IEEE*, 1999, pp. 61–67.

[9] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Trans. on Speech and Audio Proc.*, vol. 2, pp. 578–589, 1994.

[10] M.P. Cooke, *Modelling auditory processing and organisation*, Cambridge University Press, 1993.

[11] C.S. Myers and L.R. Rabiner, "A comparative study of several dynamic time-warping algorithms for connected word recognition," *The Bell System Technical Journal*, vol. 60, no. 7, pp. 1389–1409, 1981.