DETECTION OF SPIRITED INCIDENTAL MUSIC IN MOVIES

Wei-Ta Chu[1] *and Ja-Ling Wu*[2]

[1]Department of Computer Science and Information Engineering
[2]Graduate Institute of Networking and Multimedia
National Taiwan University
{wtchu, wjl}@cmlab.csie.ntu.edu.tw

ABSTRACT

This paper presents a framework for detecting spirited incidental music that is often accompanied with highlighted parts of a movie. A speech/music discrimination module is first developed to tell clips with speech or music apart. Then spirited music models are trained and used to classify the clips with incidental music into spirited or smooth ones. This two-stage framework facilitates automatic detection of spirited parts and provides a basis for efficient movie analysis.

1. INTRODUCTION

As the urgent requirements of multimedia document summarization and indexing increase, many approaches based on analyzing video and audio features have been proposed. Among them, automatic summarization of movies is still a challenging and important issue in digital content analysis research. In this paper, we focus on developing a framework that detects splendid parts, by detecting spirited background music, of movies to facilitate efficient movie analysis, such as automatic generation of movie trailers.

To enrich dramatic effects, the exciting or important movie segments often accompany spirited background music, which helps to establish the pace of scenes. Therefore, results of spirited incidental music detection provide important clues to extract important parts of movies.

Recently, many works have been done in audio classification [1] or music genre identification [2]. However, techniques about music mood detection in real-world movie applications are not widely studied. We try to develop a framework to discriminate speech and music, and identify the segments with music as impressive parts or not. In this paper, we concentrate the analysis work on action movies, where background music plays an important role to control the pace of movies and effectively provides the clues of exciting parts.

2. THE PROPOSED FRAMEWORK

A two-stage approach to detect spirited background music is proposed, as shown in Figure 1. After extracting audio features, speech/music discrimination is performed in the first stage. Based on two Gaussian mixture models, which represent the characteristics of speech and music respectively, the audio segments with background music are selected. In the second stage, spirited and smooth music models are constructed to detect spirited music. Detailed descriptions about features extraction and models constructions are described in the next few sections.

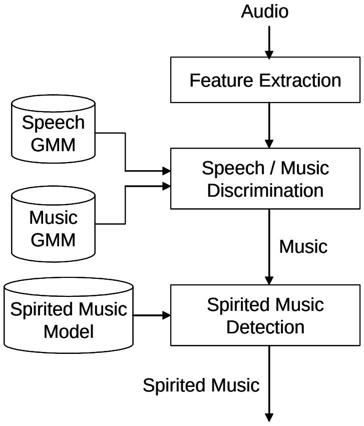

Figure 1. The proposed two-stage framework.

3. FEATURE EXTRACTION

In order to achieve promising accuracy of classification and segmentation for audio sequences, it's critical to select effective features that capture the temporal and spectral characteristics of audio signals. According to related previous works [1][3][4], some features are selected, including high zero-crossing rate ration (*HZCRR*), low short-time energy ratio (*LSTER*), mean of spectrum flux (*meanSF*), mel-frequency ceptral coefficients (*MFCC*), and entropy.

3.1. High Zero-Crossing Rate Ratio

Zero-crossing rate (*ZCR*) has been demonstrated to be useful in characterizing different audio signals. The variation of *ZCR*, which is more discriminative than the exact value of *ZCR*, is extracted in this work. *HZCRR* is defined as the ratio of the number of frames with *ZCR* larger than 1.5-fold average zero-crossing rate in a one-second (1-s) segment.

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

where n is the frame index, $ZCR(n)$ is the zero-crossing rate at the nth frame, N is the total number of frames, $avZCR$ is the average ZCR of a 1-s segment, and sgn(.) is the sign function.

Figure 2 illustrates examples of *HZCRR*s of speech and music signals. Most of the *HZCRR* values of music signal are below 0.1.

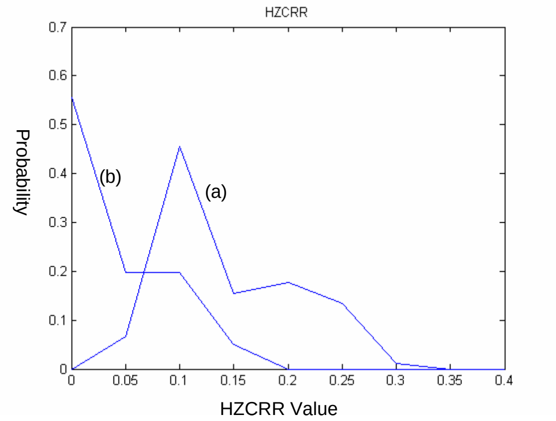

Figure 2. Probability distributions of *HZCRR*s: (a) speech and (b) music.

3.2. Low Short-Time Energy Ratio

Similar to *HZCRR*, the variation of short-time energy (*STE*) is adopted. *LSTER* is defined as the ratio of frames with *STE* less than 0.5 time of average short-time energy in a 1-s segment.

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

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 of a 1-s segment.

Figure 3 shows an example of probability distribution of *LSTER*. We can see the different characteristics of *LSTER* in speech and music signals.

3.3. Mean of Spectrum Flux

Spectrum flux (*SF*) is defined as the average variation value of spectrum between the adjacent two frames in a 1-s segment.

$$SF = \frac{1}{(N-1)(K-1)}\sum_{n=1}^{N-1}\sum_{k=1}^{K-1}[\log(A(n,k)+\delta) - \log(A(n-1,k)+\delta)]^2 \qquad (3)$$

where $A(n, k)$ is the discrete Fourier transform (DFT) of the nth frame of input signal, K is the order of DFT, N is the total number of frames and δ is a very small value to avoid calculation overflow.

We calculate the ratio of mean value of spectrum flux. As shown in Figure 4, we found that *meanSF* values of speech are higher than that of music.

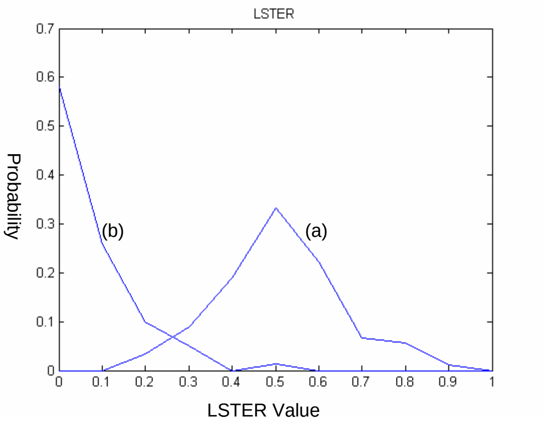

Figure 3. Probability distributions of *LSTER*s. (a) Speech and (b) music.

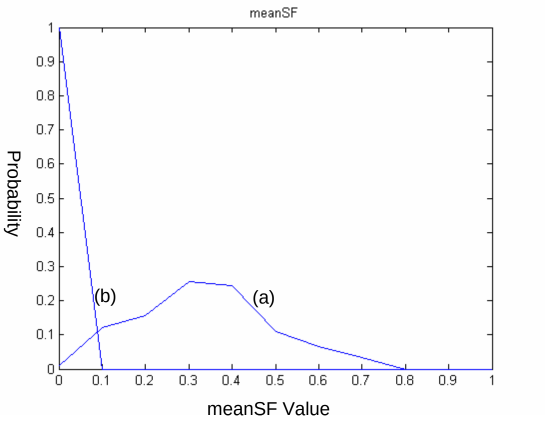

Figure 4. Probability distributions of *meanSF*s. (a) Speech and (b) music.

3.4. Mel-Frequency Ceptral Coefficients

Mel-frequency cepstral coefficients (*MFCCs*) are the most widely used features in speech recognition and other audio applications. It is computed as the inverse Fourier transform of the logarithmic spectrum in a frame. *MFCCs* are able to segment an audio clip reasonably and where the non-linear scale property of frequencies in human hearing system is considered. In this work, *MFCCs* are used in the second stage of the proposed framework to detect spirited music segments. According to our observation, we use only the first order *MFCC*.

3.5. Entropy

In comparison with smooth and spirited music, spirited music signals appear to be more disordered. To measure the 'disordered' characteristics, we evaluate a feature based on signal entropy [5], which is defined as:

$$H = \sum_{i=1}^{k} - P_i \log P_i \qquad (4)$$

where P_i is the probability of ith frame.

Figure 5 shows an example of entropy curve of a music signal. We can see the entropy value of smooth music is smaller than that of spirited signals. Samples 0 to 40 are smooth music signals, and samples 41 to 110 are spirited music signals. Therefore, we can exploit this characteristic to distinguish music characteristics.

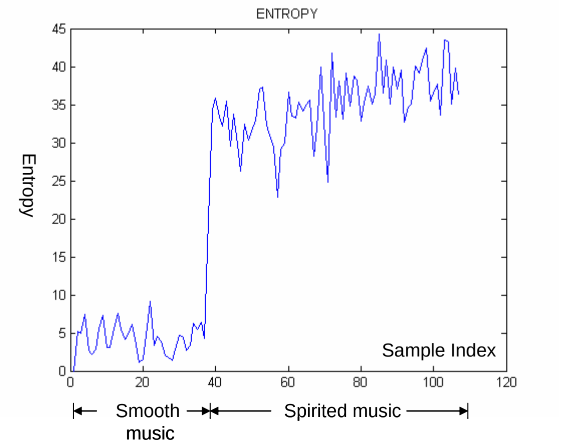

Figure 5. Entropy curve of a music signal.

4. MODELING AND DETECTION

For discriminating speech and music signals, we constructed two 3-mixtures Gaussian models based on some manually selected training data. The training data is digitized as mono-channel, 16kHz sampling frequency, with 16 bits per sample. The total length of training data for each class is about 10 minutes. The input audio signals are first split into 1-s segments with no overlapping. Then the 1-s segment is further divided into 25-ms audio frames with no overlapping, and the proposed audio features are extracted from each frame due to the stationary property of audio signals. Three features, including *HZCRR*, *LSTER*, and *meanSF* are used to characterize every 1-s segment. After extracting and concatenating features from training data, a Gaussian model with three mixtures is constructed by calculating means and standard deviations [2].

In the first stage of the proposed framework (speech/music discrimination), we define that a 1-s segment belongs to class C_i if the extracted features s = (s_1, s_2, s_3) conforms to the following criteria:

$$\begin{aligned} \text{if } & \mu_{i1} - scale * \sigma_{i1} \le s_1 \le \mu_{i1} + scale * \sigma_{i1} \text{ and} \\ & \mu_{i2} - scale * \sigma_{i2} \le s_2 \le \mu_{i2} + scale * \sigma_{i2} \text{ and} \\ & \mu_{i3} - scale * \sigma_{i3} \le s_3 \le \mu_{i3} + scale * \sigma_{i3} \end{aligned} \qquad (5)$$

where i is the class index, μ_{ij} is the mean of jth mixture of ith class model, σ_{ij} is the standard deviation of jth mixture of ith class model. In this experiment, scale is set as 1 to avoid too many false alarms.

Similar approach is also used in the second stage detection. In order to distinguish music segments into smooth and spirited parts, two Gaussian models are trained based on the features of *MFCC* and entropy. These two models are shown in Figure 6.

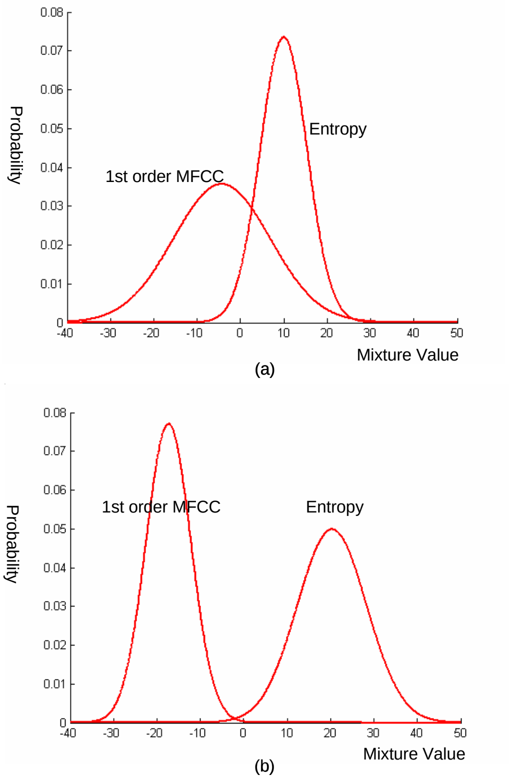

Figure 6. 2-mixture Gaussian models for (a) smooth and (b) spirited music signals.

5. EXPERIMENTAL RESULTS

Two experiments are performed to evaluate the performance of the proposed framework, including the performance of speech/music discrimination and spirited music detection.

5.1. Performance of Speech/Music Discrimination

We manually selected movie segments from several action movies to be the testing data. In order to evaluate the discrimination performance of this approach, we selected the segments which have transitions from speech or silence scenes to scenes with background music.

Figure 7 shows the discrimination result of a synthesized music signal, which contains a speech segment from 0 to 50 seconds and a music segment from 51 to 106 seconds. In the figure, the *confidence* value indicates the similarity between test data and the trained model. It is defined as:

$$confidence = 1 - abs(s - \mu_i) / scale * \sigma_i \tag{6}$$

where $abs(.)$ denotes absolute function, μ_i is the mean of ith mixture of music Gaussian model, and σ_i is the standard deviation of ith mixture of music model. The value of *scale* is the same as the setting in the training stage.

By using this approach, only a few false detections with small confidence values occurred. Actually, the spirited or splendid parts of movie segments often have continuous background music. Therefore, we can further apply some smoothing methods to filter out the parts of false detections which last for a short time duration.

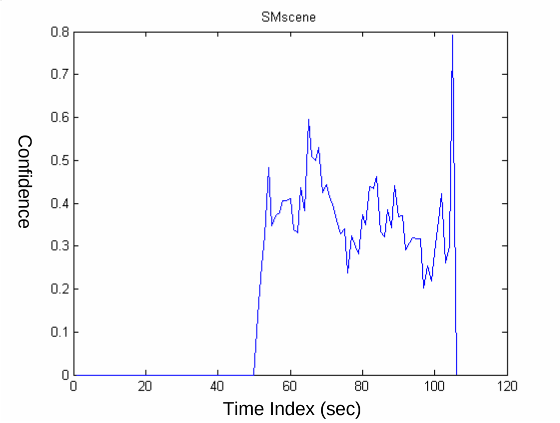

Figure 7. An example of the discrimination result.

5.2. Spirited Music Detection

After filtering out the speech segments, the remaining music segments are further classified by using the smooth and spirited music models. In the experiment, audio signals from several Hollywood movies are extracted. Figure 8 shows the results of spirited music detection for ‘matrix3.wav’, which is taken from the movie ‘The Matrix’. In this sequence, apparent spirited music starts at about 40 sec and ends at about 230 sec. We can see some false alarms on non-spirited parts. The reason for false alarms lies on the environment sounds and slight background music. For example, a role speaks from 230 to 260 sec, but the background music and sounds of thunder affects the detection results.

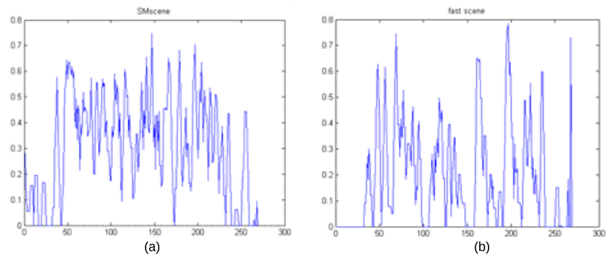

Figure 8. Results of (a) speech/music discrimination and (b) spirited music detection on ‘matrix3.wav’

6. CONCLUSION

This work focuses on detecting spirited background music in movies. A two-stage framework is proposed to discriminate speech and music and to detect spirited parts of music segments. The preliminary results show the feasibility of automatic detection of important parts of movies. Experimental results show that the proposed approach provides satisfactory results on discriminating speech and music. Although there are some false alarms, this approach performs well in most cases.

In the future, this work will be extended to automatically generate movie trailers. According to the confidence value described above, we can adaptively excerpt important segments from movies as users’ wish. Furthermore, the detection accuracy could be enhanced by taking sound effects into account or considering more audio features.

7. ACKNOWLEDGEMENT

This work was partially supported by the CIET-NTU(MOE) and National Science Council of R.O.C. under the contract No. NSC93-2622-E-002-033, NSC93-2752-E-002-006-PAE, NSC93-2213-E-002-006.

8. REFERENCES

[1] L. Lu and H.-J. Zhang, “Content Analysis for Audio Classification and Segmentation”, IEEE Transactions on Speech and Audio Processing, Vol. 10, No. 7, 2002, pp. 504-516.

[2] G. Tzanetakis and P. Cook, “Music Genre Classification of Audio Signals”, IEEE Transactions on Speech and Audio Processing, Vol. 10, No. 5, 2002, pp. 293-302.

[3] J. Pinquier, J.-L. Rous, and R. Andre-Obrecht, “Robust Speech/Music Classification in Audio Documents”, Proceedings of International Conference on Spoken Language Processing, 2002, pp. 2005-2008.

[4] S.J. Rizvi, L. Chen, and M.T. Ozsu, “MADClassifier: Content-Based Continuous Classification of Mixed Audio Data”, Technical Report CS-2002-34, 2002.

[5] T.M. Cover and J.A. Thomas, “Elements of Information Theory”, John Wiley and Sons, New York, 1991.