SPORTS AUDIO CLASSIFICATION BASED ON MFCC AND GMM

Liu Jiqing[1], Dong Yuan[1], Huang Jun[1], Zhao Xianyu[2], Wang Haila[2]

[1]Beijing University of Posts and Telecommunications, Beijing
[2]France Telecom Research & Development Center, Beijing
jiq.liu@gmail.com, yuan.dong@bupt.edu.cn

Abstract

Audio segmentation and classification can provide useful information for multimedia content analysis. In this paper, we present a approach to segment and categorize the sports audio into speech, music and other environmental sounds for sports video classification and highlight detection. We investigate the performance of Mel Frequency Cepstral Coefficients (MFCC) in a Gaussian Mixture Model frame work, and compare it to traditional short-time energy and zero-crossing rate feature. We achieve a correct identification close to 90% on MFCC with its first and second derivatives.

Keywords: audio classification; MFCC; GMM; sports audio

1 Introduction

In recent years, with the boom of audiovisual content on the internet, multimedia information retrieval has become a prevalent technology, which has created many important applications in professional media production, audiovisual archive management, education, entertainment, surveillance, and so on. Audio data is an important part of many modern multimedia and computer applications. The audio signal may actually play a primary role in content parsing of audiovisual data, for example, the speech information contained in audio signals is usually critical in identifying the theme of the video segment. As the volume of the available audio content increases dramatically, manual segmentation and indexing becomes impossible. Automatic segmentation and classification through computer processing based on audio content analysis is clearly the trend.

Actually, over many years, a lot of efforts have been made to develop methods for extracting information from audio data. John Saunders [1] presented a straightforward approach to classify speech and music in radio broadcasts application based on the statistics of the energy contour and the zero-crossing rate. Eric Scheirer report a speech/music discriminator based on various combinations of 13 features, such as 4 Hz modulation energy, spectral centroid, and zero-crossing rate [2]. Several classification strategies, including Gaussian mixture models and K-nearest-neighbor classifiers were evaluated. Lie Lu and Stan Z. Li [3] also used a combination of spectrum-based features and perceptual-based features for general audio classification, with prevalent SVM as classifier. In Ajmera's work [4], entropy and dynamism features based on posterior probabilities of speech phonetic classes are used to form an observation vector sequence, which is used in a HMM classification framework.

In our work, in order to provide information for sports video classification, we segment the sports audio to 2-seconds audio clips and categorize them into speech, music and other environmental sounds. Considering sports commentary takes up main durations, we choose Mel Frequency Cepstral Coefficients (MFCC) as our major audio feature, which have been proved to be representative on speech in our former work on Text-Independent Speaker Verification [5][6]. We also carry out a comparison with short-time energy and zero-crossing rate in the Gaussian mixture model classifier frame work.

The rest of the paper is organized as follows. Section 2 discusses in detail the audio features used in audio classification. Section 3 presents the audio segmentation and classification scheme. In Section 4, experiments and performance evaluation of the proposed algorithms are presented.

2 Audio Features Analysis

Discriminative features will contribute a lot in audio classification task. In order to improve the accuracy of classification and segmentation for audio sequence, it's important to choose the features that can represent the temporal and spectral characteristics properly. In our system, we select the mel frequency cepstral coefficients (MFCC), which was proved to be effective for speech and music discrimination [7]. Also, as a comparison, we extract the traditional short-time energy and zero-crossing rate.

978-1-4244-4591-2/09/$25.00 ©2009 IEEE

2.1 Mel-frequency Cepstral Coefficients

Mel-frequency cepstral coefficients (MFCCs) have been commonly used in speech recognition. They are based on the known variation of the human ear's critical bandwidths with frequency. The MFCC technique makes use of two types of filter, namely, linearly spaced filters and logarithmically spaced filters. To capture the phonetically important characteristics, signal is expressed in the Mel frequency scale, which is shown in Figure 1.

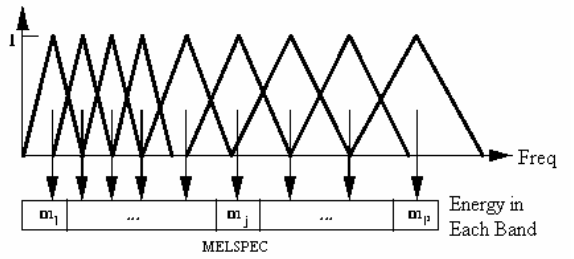

Figure 1. Mel-Scale Filter Bank

The calculation of MFCC can be divided into 5 steps, which are shown as follows:
1. Divide signal into frames.
2. For each frame, take the Fourier transform.
3. Take the logarithm.
4. Convert to Mel (a perceptually-based) spectrum.
5. Take the discrete cosine transform (DCT).

In the 5th step, the DCT are defined as:

$$c_j = \sqrt{\frac{2}{N}} \sum m_j \cos\left(\frac{\pi i}{N}(j-0.5)\right), j = 1, \cdots, L \quad (1)$$

Where N is the number of filterbank channels, and L is the desired length of the cepstrum. Here, we set N=24 and L=12.

In our implementation, The FFT should use a Hamming window and the signal should have first order pre-emphasis applied using a coefficient of 0.97. 12 MFCC coefficients plus C0 are computed, the frame period is 10 ms, and the window size is 25 ms. And to represent the dynamic information of the feature, we compute the first and second derivatives, and append them to original feature vector to form a 39-dimension feature.

2.2 STE and ZCR

The short-time energy (STE) and zero-crossing rate (ZCR) are the basic features for speech recognition. STE reflects the representation of the amplitude variation over time, while ZCR provides a measure of the weighted average of the spectral energy distribution in the waveform, the spectral center of mass. John Saunders presented that they are useful in speech/music discrimination [1] based on an heuristic method.

In our system, we compute the STE and ZCR for each frame, and then combine them to a feature vector as the input of GMM models. The implement in detail is explained in [8].

3 Segmentation and Classification

3.1 Segmentation

Before classification is carried out, the long-term audio sequence should be cut into short audio segments. In our scheme, the audio signal is assumed to be monophonic. In the case of multi-channel audio signals, the average value per-sample across multiple channels is taken as input. All input signals are down-sampled into 16-KHz sample rate and subsequently segmented into sub-segments by 2-s window, without overlap. This 2-s audio clip is taken as the basic classification unit in our algorithms. If there are two audio types in 2-s audio clip, it will be classified as the dominant audio type. For each audio clip, features described above is extracted every 10ms, with a 15ms overlap, i.e. a feature vector is extracted over a 25ms window.

3.2 Classification

The classification for each audio clip is carried out using Gaussian Mixture Models (GMM) [9].

For every D-dimension feature vector $\vec{x}$, its likelihood to a claimed audio type model $\lambda = \{p_i, \vec{u}_i, \sum_i\}, i = 1, ..., M$ is defined as follows:

$$p(\vec{x} \mid \lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \quad (2)$$

In that, $b_i(\vec{x})$ is a Gaussian function:

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2} |\sum_i|^{1/2}} \exp\{-\frac{1}{2}(\vec{x}-\vec{u}_i)^T \sum_i^{-1} (\vec{x}-\vec{u}_i)\} \quad (3)$$

Each of the possible classes of signal, are represented by a GMM trained on the training set using the expectation maximization (EM) algorithm. The variances of the distributions were modeled by a diagonal covariance matrix. Tests were carried out on every frame to get a probability score, and the overall score summation for each audio clip will decide the sound type of this segment.

4 Experiments on Sports Audio

4.1 Data Sets

Unlike the data-set used in the general work, our database of sports audio is not composed of relatively clean audio such as CD recordings and TIMIT database. It is from broadcast TV which is an un-controlled audio environment with high background interference that makes the audio classification more difficult.

We've collected over 10 hours audio evaluation data from TV broadcasting of soccer, tennis, car racing and skee. Each of them is hand-labeled into one of the three classes as ground truth: speech, music and other environmental sounds, In which, other sounds class is composed of applause, cheer, motor and so on. The respective durations for 3 classes are: speech 483 minutes, music 54 minutes and other sounds 77 minutes. We can see that speech holds a majority in sports audio.

4.2 Evaluation Measure

In the experiments, we adopt the common used F-measure as our evaluation measure, which is the harmonic mean of the precision P and recall R defined as:

$$F = \frac{2PR}{P+R} \tag{4}$$

Where, take speech for example:

$$P = \frac{\text{duration of correctly detected as speech}}{\text{total detected speech duration}} \tag{5}$$

$$R = \frac{\text{duration of correctly detected as speech}}{\text{total duration of groundtruth speech}} \tag{6}$$

4.3 Results

On the evaluation data, we first explore the performance of different features with 128 mixtures GMM as classifier. Then we discuss the influence of different GMM model order, that will answer why we choose 128 in feature comparison.

A. Feature Comparison

For cepstral features in the experiments, 12 MFCC coefficients plus C0 are computed. First and second order derivatives computed over 5 frames are appended to each feature vector, which results in dimensionality 39, we mark it as MFCC_0_D_A. The results are shown in Table 1. In that, the overall performance means, precision and recall is calculated on all 3 sound types by durations. As the majority of speech, the performance of speech tends to dominate the overall performance.

Table 1 Results on MFCC_0_D_A

Audio Type	Recall	Precision	F-score
Speech	0.9432	0.7845	0.8565
Music	0.3796	0.4878	0.4269
Other Sounds	0.8586	0.8793	0.8688
Overall	0.8021	0.8021	0.8021

As a comparison, we extract the STE and ZCR features, and combine them to a vector for each frame. Table 2 doesn't show a remarkable performance as MFCC. That is because it doesn't represent the characteristics properly like MFCC, and noisy environment maybe harm the discrimination of zero-crossing rates.

Table 2 Results on STE&ZCR

Audio Type	Recall	Precision	F-score
Speech	0.8723	0.7235	0.7909
Music	0.3606	0.4318	0.3930
Other Sounds	0.5941	0.2742	0.3752
Overall	0.7232	0.7232	0.7232

Considering that the sports audio is recorded with high background noise. To alleviate the interference, we carried out a Cepstral Mean Normalization (CMN) on original 13-dimension MFCC before computing the first and second order derivatives, which is a very effective technique in practice where it compensates for long-term spectral effects such as those caused by different microphones and audio channels. We mark it as MFCC_0_Z_D_A, in that, Z means CMN. Table 3 shows that the CMN indeed enhance the performance in a large scale.

Table 3 Results on MFCC_0_Z_D_A

Audio Type	Recall	Precision	F-score
Speech	0.9745	0.9611	0.9677
Music	0.6253	0.7059	0.6631
Other Sounds	0.8620	0.8050	0.8325
Overall	0.8896	0.8896	0.8896

From above 3 tables we see that, all features give better performance on speech than non-speech, which may be caused by following reasons. First, the features were designed for speech or speaker recognition, so they can represent the characteristic properly while don't manifest much discrimination between music and other environmental sounds. Second, music and environmental sounds are more similar in the spectrum, while speech tend to be band-limiting (mainly between 500~3000Hz). Besides, some elements of music, e.g. drumbeat, are noise-like in acoustics.

B. GMM Model Order Selection

Determining the number of components M in a mixture needed to model an audio type is an important but difficult problem. There is no theoretical way to estimate the number of mixture components a priori. To investigate the audio classification performance of the GMM with respect to the number of component densities per model, the following experiment was conducted on 16, 32, 64, 128, 256, 512 component Gaussian densities. Figure 2 shows the F-score versus the number of Gaussian components.

In Figure 2, we find that the system performance will increase with the GMM model order, it's intuitive that the increase of model capacity could depict more details of the audio signals. Meanwhile, we can see the curve becomes gentle after the model has

contained 128 Gaussian mixtures. Considering the balance between system performance and model complexity, we choose 128 mixtures when training a sound type model, like situations in the above experiments.

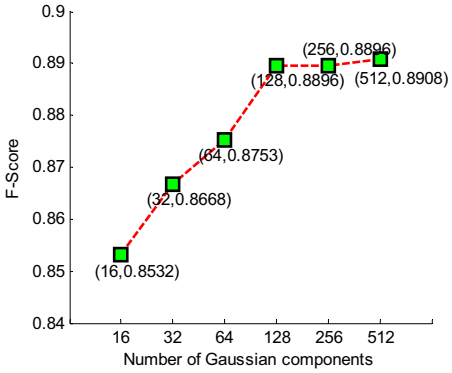

Figure 2. The system performance versus the number of Gaussian components.

5 Discussion

In this paper, we have proposed a scheme for audio segmentation and classification as speech, music or environmental sounds on the sports audio data. Our algorithm is discriminative between speech and non-speech, while it perform not so well relatively between music and other environmental sound. More descriptive features between music and other noise should be considered in the next step. And more knowledge on characteristics of different types of sports audio data should be investigated, which will provide more information for fine audio classification, sports classification and sports highlight detection.

In the future, our work on audio classification could be extended to a more detailed categorization and description of audio. They could be the first level of a hierarchical classifier, and then continue to classify the audio data to more specific categories, such as speaker segmentation, music genre classification, motor and ball-hit sound detection, and so on. Meanwhile, the audio classification results could be combined with content-based indexing of visual data. This will provide an effective method of multimedia information retrieval, which could be used in many applications we introduced in the introduction section.

Acknowledgments

Supported by The Key Project of The Ministry of Education of P. R. China (108012), and Scientific Research Fund of Overseas Returned Staff, Ministry of Education.

References

[1] J. Saunders, "Real-time discrimination of broadcast speech/music," *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, 1996, pp. 993-996 vol. 2.

[2] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, 1997, pp. 1331-1334 vol.2.

[3] Lie Lu, Hong-Jiang Zhang, and Hao Jiang, "Content analysis for audio classification and segmentation," *Speech and Audio Processing, IEEE Transactions on*, vol. 10, 2002, pp. 504-516.

[4] J. Ajmera, I. Mccowan, and H. Bourlard, "Speech/music segmentation using entropy and dynamism features in a HMM classification framework," *Speech Communication*, vol. 40, 2003, pp. 351-363.

[5] Yuan Dong, JIAN ZHAO, Liang Lu, Jiqing Lui, Xianyu Zhao, and Haila Wang, "Eigenchannel Compensation and Symmetric Score for Robust Text-Independent Speaker Verification," *Chinese Spoken Language Processing, 2008. ISCSLP '08. 6th International Symposium on*, 2008, pp. 1-4.

[6] Y. Dong, L. Lu, X. Zhao, and J. Zhao, "Studies on Model Distance Normalization Approach in Text-Independent Speaker Verification," 2009, pp. 556-560.

[7] M. Carey, E. Parris, and H. Lloyd-Thomas, "A comparison of features for speech, music discrimination," *Acoustics, Speech, and Signal Processing, 1999. ICASSP '99. Proceedings., 1999 IEEE International Conference on*, 1999, pp. 149-152 vol.1.

[8] Y. Dong, L. Lu, and H. Wang, "Confusion based automatic question generation," *IET Conference Publications*, vol. 2008, Jan. 2008, pp. 64-67.

[9] D.A. Reynolds and R.C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE transactions on Speech and Audio Processing*, vol. 3, 1995, pp. 72-83.