Detecting sound events in basketball video archive

Dongqing Zhang, zd35@columbia.edu
Dan Ellis, dpwe@ee.columbia.edu

Electrical Engineering Department of Columbia University
New York City, NY 10025

Abstract

The report proposes a method for detecting the sound events in a basketball game with focusing on detecting cheering sound. MFCC (Mel-frequency cepstral coefficient) features are used to identify the cheering sounds from speeches and other confusing sounds. The mfcc features are fed into a neural network and classified into three classes (cheering, speech, and others). To improve the MFCC-NN performance, a measure for temporal spectral variation is proposed, which is defined by LPC coefficient entropy. Normalized energy is also used to eliminate those false alarms caused by background noise. The outputs from these three channels are finally fused together and postprocessing techniques are used in order to get robust results. For other events, such as dribbling, template matching based approach is proposed. Experiments showed our methods achieved good performance for very difficult sound track. The described method can be used in basketball video content retrieval and highlight extraction.

Keywords
Basketball, Highlights, *mfcc*, LPC, events detection

1. INTRODUCTION

With the explosion of online video data, video content management has become a very hot research topic in these years. Many interesting topics have been extensively investigated, these includes video segmentation, summarization, highlight detection etc [1].

For the application like online streaming, and large video archive management, the data is very huge, and network bandwidth is very limited. Detecting video highlight events become very useful for those applications. For example, in sports video, the video editor need to extract the meaningful events from a large archive, and the video website need only broadcast those highlighting stories because of the network bandwidth limitation.

Previous researchers have investigated highlight and event extraction for sports video. These include soccer video, baseball video, etc [2][3][4]. All of these video data belong to the field sport video. Another category is court video, including basketball, badminton, etc. Because of being held inside building, they have different visual and audio

characteristic from the field sports video. Court sports game has particular lighting condition, and rich game specific sound events.

In this research project, we attempt to explore the automatic approach for analyzing the basketball game video achieves, a very important category of court sports game. Basketball game is very popular in North American. There are huge baseball video data because of the long history of basketball game. With the accumulation of these video data, the management issue has been more and more important. In contrast to many previous video analysis technology which employ visual features to do video segmentation, we use audio data, which is more efficient than visual feature analysis.

We classify the sounds in a basketball game into two categories: Generic sound events, which are common for different sports game, including cheering, reporter voice, and whistle; basketball game specific sound, including shooting, ball bouncing from board, dribbling. Although generic sound events are very similar in different sports game, in basketball game, it is more difficult to be detected due to more confusing sounds. To those specific sounds, we found it is very challenging to detect them. However, since the occurrence of these sounds are highly correlated with the occurrence of audience cheering. We can infer the location of these sounds by detecting the audience cheering.

To cheering sounds, we propose a framework that combines both spectral and temporal features. In spectral domain, *Mfcc* and their derivatives are used to characterize the spectral properties. These features are extracted from a small window, and are fed into a Neural Network to do initial classification. Each window is tagged by a class label after classification. These classes are cheering, speech, and other sounds. Neural network is trained on samples that are obtained by manual classification. In temporal domain, we observed that for cheering sound, the spectral features are quite stable and much less variable comparing with speech and other sounds. To represent the temporal stationary, entropy measure is developed based on LPC coefficients. Since cheering sound is relative louder than other sounds, normalized energy is incorporated to enhance the detection performance. Finally the results from these three channels are combined together to obtain classification results. For the sake of robustness, hysteresis thresholding and morphological filter are employed, such that the system can be immune to the noise produced by NN and thresholding.

To detect the game specific events like cheering and dribbling, we employed a temporal matching method, however the experiments showed that the temporal matching method is very unstable in the difficult sound track.

2. RELATED WORK

There are various approaches to track the highlight in a video: use visual features, such as histogram, motion, color, etc; analyze by audio; and analyze by artificial meta data, like closed caption, text on video. Among these features, visual feature analysis needs lots of computation resources, especially for motion features. Comparatively, audio analysis is more efficient and thus has attracted many researchers.

Sports video analysis has been an interesting topic in video indexing and retrieval. Gong et. Al. [1] studied parsing TV soccer program, they analyze the video data by tracking and detecting the positions of players using motion vector. Babaguchi, et. Al [4]

Proposed an event based video indexing method, they take the idea of intermodal collaboration to combine multimodal information streams, which include visual, auditory and text. Chang, et al [3]. explored a speech analysis method for video indexing. They use keyword spotting and cheering detection to locate the meaningful segment of video. Afterwards they use video analysis to analyze the soccer video. Rui,L, et al [5] developed a system for extracting highlights from TV baseball program. They employ speech endpoint detection using phoneme-level features, and use reporter's exciting speech to infer the occurred important events. A template matching based method is employed to detect the specific events.

This reporter is targeted at basketball video analysis using audio features, different from Rui's method, we use cheering detection instead of speech detection. We employ a hybrid method to incorporate both spectral and temporal features.

3. CHEERING DETECTION

Cheering is the salient sound event in a basketball game, it occurs always after or before an exciting event. So cheering detection is very important for other event tracking. Cheering has its distinct characteristics: (1) No phoneme structure is present in the spectrum of cheering sound. (2) In time domain, cheering usually last a quite long time, which makes the spectrum stationary in time domain (3) Cheering is normally louder than other sounds, which means sound energy can be used to detect cheering. Based on these observations, we employ *mfcc* features to characterize spectral properties, and the entropy of LPC coefficients in a large time window for temporal characteristics. The system can be illustrated as the following flowchart:

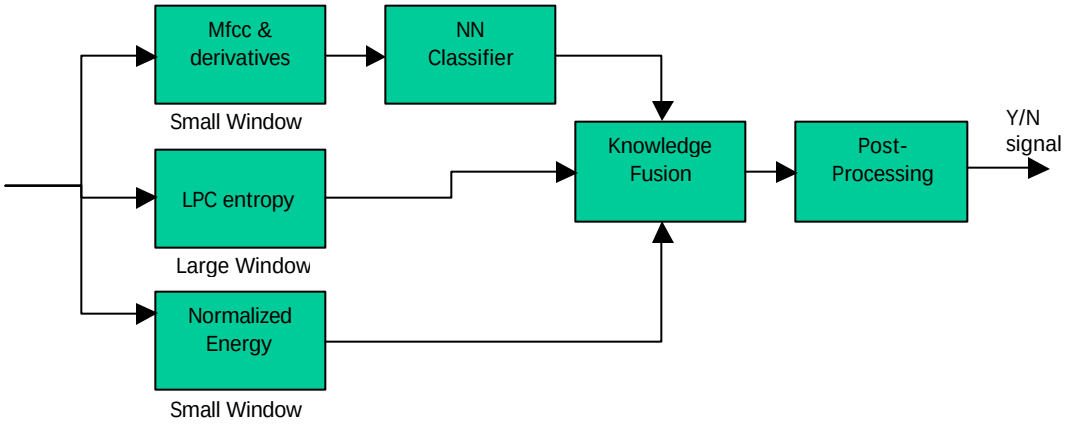

Fig.1 System Flowcharts

In the workflow, Neural network is used to classify the sound window based on *mfcc* features, knowledge fusion model combines the signals from NN, LPC entropy, and normalized energy to get a binary signal, where 1 means cheering sound, while 0 means

not. To eliminate noise, a hysteresis thresholding algorithm is used. The morphological filter is introduced to get rid of the noise produced by NN.

3.1 Mfcc features

Mfcc (Mel-scale cepstral coefficients) features have been successfully used in speech. Different from the traditional cepstral coefficients, *mfcc* computes the cepstral on the warped mel-scale spectrum, which has finer resolution at the low frequency while coarser resolution at high frequency. The mel-scale spectrum zooms into the "phoneme area" of the spectrum of a segment of speech. Various features defined on Mel-scale spectrum has been developed for speech detection and recognition, *mfcc* is one of these features, which has demonstrated its advantage against other features.

In order to detect the cheering events, we have to discriminate cheering from speech and other sounds. Since *mfcc* can well modal the characteristic of speech spectrum features, it should be effective for modeling the cheering spectrum. 13 *mfcc* features are used in this system, *Mfcc* features are extracted at 20 ms window, or 50 frames per second.

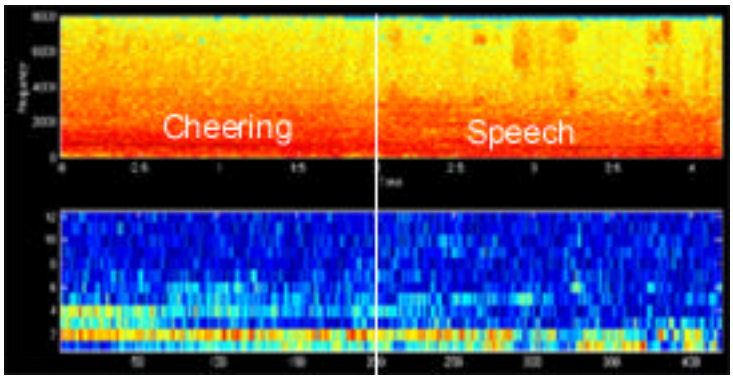

Fig 2. *Mfcc* features of cheering and speech

Besides *mfcc*, *mfcc* derivative is also useful for speech recognition. *Mfcc* derivatives are able to capture the spectrum relationship between the adjacent frames. Thus they can also be used to identify cheering, since for cheering sound, spectrum is almost smooth, while in speech it is not. In experiments, we found that neural network classification performance increases 1-2 percent by introducing *mfcc* derivatives. This indicates the effectiveness of *mfcc* derivative features.

3.2 Neural Network classification

Neural network is applied to do initial classification of *mfcc* windows. Neural network is a kind of non-linear classifier, which is desirable for the classification in high dimensional and nonlinear feature space. The *mfcc* windows are classified into three

classes: cheering sound, speech sound and other sound. The past experience of Neural network has demonstrate that even for a two class application scenario, it is better to use three classes rather than two. Using three classes is able to avoid overclassfication. The "other sound" class actually takes this task in the system. Although they are frequently confused with speech class, they rarely interfere with the cheering class.

Neural network take 10 *mfcc* coefficients and 5 *mfcc* derivatives as its input, and produce three outputs, the class labeling is based on the three outputs. Basically it takes the index of the maximum output as the class label. The neural network has 5 nodes in its hidden layer. We manually collect various samples for those three classes for neural network training, 88 samples for each classe, each sample takes one *mfcc* window.

The following is the neural network result for a dunk shot segment. The cheering sound locates in the right side of the white line. Notice that although speech is easily confused with other sounds, cheering rarely does.

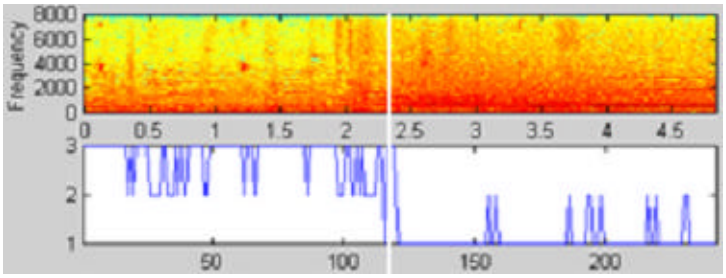

Figure 3. Neural Network classification results

3.3 LPC entropy as temporal feature

Mfcc features only take advantage of the spectrum feature in a small window. However looking at the large windows of a sound track, we find that the spectrum of cheering sound is almost constant. This property is distinct for cheering sound and does not exist in speech sound. This indicates that cheering detection might be better than speech endpoint detection for exciting highlight detection.

In [5], Rui, Y used an entropy feature defined on the spectral components. Similarly, we also use entropy to measure the spectral variations in time domain. Here, entropy is only a measure for signal stationary instead of information coding cost. By comparing spectrum with LPC coefficients, we found that LPC coefficients map is more stable than signal spectrum. The reason might be that LPC coefficients are a kind of polynomial approximation for signal spectrum envelop, thus effectively eliminate the effect of noise. Therefore, LPC coefficients are more suited to define the entropy measure.

We define the signal stationary measure in time domain as the average entropy of each LPC component:

$$Etr = -\frac{1}{D}\sum_{d=1}^{D}\sum_{n=1}^{w} P_{dn}\log P_{dn}$$
$$P_{dn} = |a(n,d)|^2 \Big/ \sum_{n=1}^{w} |a(n,d)|^2 \qquad (1)$$

Where *a* is the LPC coefficients, *w* is the time window size, we use 15 LPC coefficient frames in a window, where each LPC coefficient is extracted in 512 audio samples. *D* is the LPC order, with 8 in experiment. The equation obtains its maximum value when each LPC component is invariant on the large time window. For LPC with 8 components, the maximum value is 2.708.

The following figure shows the LPC coefficients and the entropy measure of a segment of sound, which contains cheering at left and speech at right.

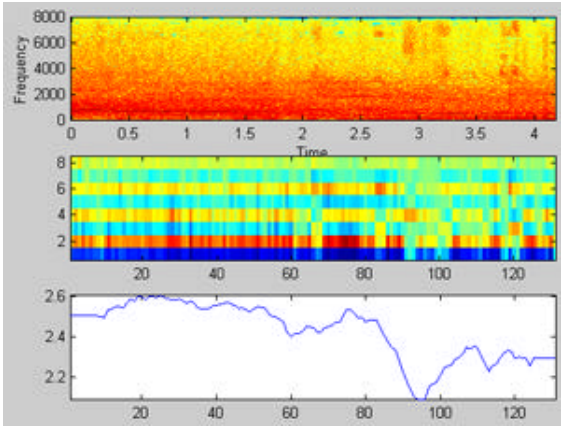

Figure 4. LPC coefficients and LPC entropy

In the figure, notice that the entropy measure of cheering is approximate to the maximum value. And speech has lower value because of the variation of LPC coefficients.

3.4 Normalized energy and post-processing

Mfcc and LPC entropy does very well to discriminate the cheering, speech and some specific sounds. However it encountered difficulties to discriminate background noise and cheering, since background noise have similar spectrum shape with cheering. To deal with this problem, normalized energy is introduced to classify these two sounds. This is based on the fact that cheering sound is much louder than background noise.

Information fusion is needed to combine these three outputs, there are two possible schemes, we call them "hard" fusion and "soft" fusion. In "hard" fusion, all signal are binarized into true/false signal, and "AND" operator is used to combine these binary signals. This scheme would bring with the noise because of NN classification and thresholding. Postprocessing is needed for this scheme. Another scheme is "soft" fusion, which uses a statistical framework to combine the signal from different channels. The idea is, when certain channel is uncertain to classify the signal it put the weight to the other channels. This scheme should be better than "hard" fusion.

For simplify our system, only "hard" fusion is used. The fusion output is 1 if the sound window is cheering, otherwise 0. However, in order to eliminate the noise, two techniques are employed, one is hysteresis thresholding [6], which employs two thresholds to binarize the signal, the selective threshold is used to get the initial '1' windows, and the expanding threshold is used to extend the '1' windows to their

neighborhood area until the signal value falls below the second threshold. The other technique is morphological filter [7], which fill the gap caused by narrow '0' window, and throw away those cheering segments with too short duration.

The figure below displays the filtering result of morphological filter.

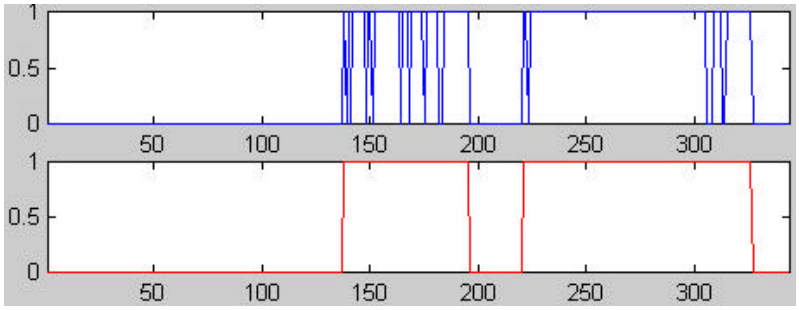

Fig 5. Morphological filter to eliminate the noise

4. SPECIFIC EVENT DETECTION

Game specific event is more difficult to be detected than cheering, the reason is they are usually of very short duration, and more vulnerable to the interference by other sounds. Furthermore, some sounds are even similar to each other, such as ball hitting the basketball board and ball bouncing.

We attempted to use template matching method to detect the specific sound like dribbling. The initial test results showed that template matching is not very good at detecting the dribbling sound. Since the duration of dribbling sound is very short, we use a template only with 6 spectrum window to match the spectrum of the sound track, it works well on sound segment with clear background, but get inaccurate results on those sounds segment with noisy background.

Obviously, use one fixed event template is hard to obtain good performance for those hard sound tracks. More effective approach is to use a group of orthogonal template which expand a small dimensional subspace for specific sound event. The matching can be achieved to compute the distance from the expanded subspace. This can be realized by Principle Component Analysis (PCA).

5. EXPERIMENTS

We use a very hard sound track captured form NBA.com website, which is raw sound track captured in the basketball court before any processing. The reason to use this sound track is because the real useful application of the analysis algorithm is for those raw video/audio data before editing and processing, such that the system can help the video editor to locate, extract and manage the meaningful part of a video file. The length of the experimental sound track is about 4 minutes.

5.1 Evaluation metrics and results

Evaluation metrics is defined depending on different applications. For example of the cheering event detection, we can measure the performance by counting the error windows. However for highlight extraction, we may don't need such high detection granularity. For most application, hitting cheering event is enough and the accurate location is secondary concern. Thus, we define two kinds of evaluation metrics: window hitting error metric(WHR), and event hitting error metric (HER). These measures include two kinds of error metrics: false acceptance rate and false rejection rate. False detection means the system recognizes a non-cheering sound window as cheering or non-cheering event as cheering. False rejection is vice versa.

The Results are listed in the following table.

Table 1. Valuation results

	False acceptance	False rejection
Window hitting	234/953 = 24.6%	309/10067 = 3.06%
Event hitting	4	0/9

In event hitting, 4 means there's four false acceptances for all data, and there's no false rejection for 9 cheering events.

5.2 Discussion

We achieve a good false rejection rate for event hitting, which means the system can detect every cheering event occurred. However there are also a few of false acceptance. Having observed the false acceptance samples, we found most of these errors occurred because of Neural Network output, which falsely classifies non-cheering window as cheering window. This could be solved by training NN using more samples. In particular, use the error window to force the NN change its decision surface.

6. CONCLUSION

We have proposed a scheme for detecting the cheering event in a basketball game, which is used to analyze the basekball game video. The method combine *mfcc* features, LPC entropy and normalized energy. Post-processing techniques are used to eliminate noise. We also attempted to use template matching to detect the specific sound event. Our experiments showed that the method perform well in a difficult sound track. Which demonstrates the combination of *mfcc* features and LPC entropy is very effective to identify the cheering sound in basketball game. Further work would cover advanced specific event detection using PCA and more robust training scheme for Neural Network.

7. REFERENCE

[1] Zhang, H., et al. Automatic Parsing of News Video. In IEEE Conference on Multimedia Computing and Systems. 1994.

[2] Gong, Y. et al. Automatic Parsing of TV Soccer Programs. In IEEE Conf on Multimedia Computing and Systems, 1995.

[3] Chang, Y.L. et al. Integrated Image and Speech Analysis for Content-based Video Indexing. In IEEE Conf. On Multimedia Systems and Computing. 1996.

[4] Noboru BABAGUCHI, Yoshihiko KAWAI and Tadahiro KITAHASHI: Event Based Video Indexing by Intermodal Collaboration, Proc. First International Workshop on Multimedia Intelligent Storage and Retrieval Management (MISRM'99) in conjunction with ACM Multimedia Conference 1999, Orlando, pp.1-9(1999-10).

[5] Automatically Extracting Highlights for TV Baseball Programs, Proceeding ACM Multimedia 2000, p105-115.

[6] Canny, A.. A computational approach to edge detection. *IEEE Trans. PAMI*, 8:769 - 698. (1986).

[7] Fundamentals of Digital Image Processing - A.K. Jain. (1988).