AUDIO EVENTS DETECTION BASED HIGHLIGHTS EXTRACTION FROM BASEBALL, GOLF AND SOCCER GAMES IN A UNIFIED FRAMEWORK

Ziyou Xiong†, Regunathan Radhakrishnan‡, Ajay Divakaran‡ and Thomas S. Huang†

†Department of Electrical and Computer Engineering,
University of Illinois at Urbana-Champaign,
Urbana, IL 61801
‡Mitsubishi Electric Research Laboratory at Murray Hill,
Murray Hill, NJ 07974
E-mail: {zxiong, huang}@ifp.uiuc.edu, {regu, ajayd}@merl.com

ABSTRACT

We developed a unified framework to extract highlights from three sports: baseball, golf and soccer by detecting some of the common audio events that are directly indicative of highlights. We used MPEG-7 audio features and entropic prior Hidden Markov Models(HMM) as the audio features and classifier respectively to recognize these common audio events. Together with pre- and post-processing techniques using general sports knowledge, we have been able to generate promising results dealing with the audio track that is dominated by audio mixtures and noisy background.

keywords: Sport Highlights, Unified Framework, Audio Features, HMM

1. INTRODUCTION AND RELATED WORK

Most of the current systems focus on a particular sport when highlights are extracted. For baseball, Rui et al[1] have detected the announcer's excited speech and ball-bat impact sound using directional template matching based on the audio signal only. For golf, Hsu[2] has used Mel-scale Frequency Cepstrum Coefficients(MFCC) as audio features and multi-variate Gaussian as classifier to detect golf club-ball impact. For soccer, Xie et al[3] and Xu et al[4] have proposed to segment soccer videos into play and break segments using dominant color and motion information.

In this paper we explore the possibility to build a unified framework to extract highlights from all these three sports. Our motivation stems from computational power constraints on set-top devices such as TiVo and WebTv as well as on personal digital video recorder applications. Such constraints, especially on hardware devices, rule out having a completely distinct highlights extraction algorithm for each sport, and thus motivate us to look for general features that would work across different sports. We are currently concentrating on audio since audio processing is computationally simple and lends itself better to extraction of content semantics.

In Section 2, we describe some observations regarding common events across different sports. These observations enable us to choose good audio features and classifiers to recognize these events, which we describe in Section 3. We propose our approach and experimental results in Section 4 and Section 5 respectively. We conclude in Section 6.

2. COMMON EVENTS ACROSS DIFFERENT SPORTS

In the audio domain, there are common events relating to highlights across different sports. After an interesting golf hit or baseball hit or an exciting soccer attack, the audience shows appreciation by applauding or even loud cheering. The duration of applause or cheering is longer when the play is more interesting(e.g, a home-run in baseball). There are also common events relating to un-interesting segments in sports TV broadcasting, e.g, TV commercials that are mainly composed of music or speech with music segments.

Our observation is that the audience's applause or cheering are more general across different sports than the announcers' excited speech. We hence look for robust audio features and classifiers to classify and recognize the following audio signals: applause, cheering, music, speech, speech with music. The former two are used for highlights extraction and the latter three are used to filter out the un-interesting segments.

3. AUDIO FEATURES AND CLASSIFIER

We have studied and compared one of the widely used audio features, i.e, MFCC[5] and the newly adopted MPEG-7 audio features[6]. We have also compared Entropic Prior

0-7803-7663-3/03/$17.00 ©2003 IEEE

	[1]	[2]	[3]	[4]	[5]	[6]
[1]	1.00	0	0	0	0	0
[2]	0	0.923	0	0	0.077	0
[3]	0.125	0	0.875	0	0	0
[4]	0	0	0	0.944	0.056	0
[5]	0	0	0	0	0.941	0.059
[6]	0	0	0	0	0	1
	Average Recognition Rate: 94.728%					

Table 1. Recognition Matrix on a 90%/10% training/testing split of a data set composed of 6 classes. [1]: Applause; [2]: Ball-Hit; [3]: Cheering; [4] Music; [5] Speech; [6] Speech with Music. The results here are based on MPEG-7 Audio Features and EP-HMM with trimming of states and model parameters.

HMM(EP-HMM)[7] with the traditional Maximum Likelihood HMM(ML-HMM) for classification purpose. In [8], we report that for our quite noisy sports audio database, on average the best combination is MPEG-7 features with EP-HMM with trimming of states and model parameters out of 6 different feature-classifier pairs. Its recognition matrix is shown as in Table 1.

In the following, we give a brief introduction to MPEG-7 features and EP-HMM. For a detailed explanation please see [8].

3.1. MPEG-7 Audio Features

The MPEG-7 features consist of dimension-reduced spectral vectors obtained using a linear transformation of a spectrogram. They are the basis projection features based on Principal Component Analysis(PCA) and an *optional* Independent Component Analysis(ICA). For each audio class, PCA is performed on the normalized log subband energy of all the audio frames from all the training examples in the class. The frequency bands are decided using the logarithmic scale(e.g. an octave scale).

3.2. Entropic Prior HMM[7]

Denote λ as the model parameters, O as the observation. When we don't have any bias toward any prior model λ_i, that is we assume $P(\lambda_i) = P(\lambda_j)$, $\forall i, j$, then the Maximize A Posteriori(MAP) test is equivalent to the Maximum Likelihood(ML) test: O is classified to be of class j if $P(O|\lambda_j) \geq P(O|\lambda_i)$, $\forall i$ due to the Bayes rule: $P(\lambda|O) = \frac{P(O|\lambda)P(\lambda)}{P(O)}$. However, if we assume the following biased probabilistic model $P(\lambda|O) = \frac{P(O|\lambda)P_e(\lambda)}{P(O)}$, where $P_e(\lambda) = e^{-H(P(\lambda))}$ and H denotes entropy, i.e, the smaller the entropy, the more likely the parameter, then we must use the MAP test and compare $\frac{P(O|\lambda_i)e^{-H(P(\lambda_i))}}{P(O|\lambda_j)e^{-H(P(\lambda_j))}}$ with 1 to see whether O should be classified to be of class i or j.

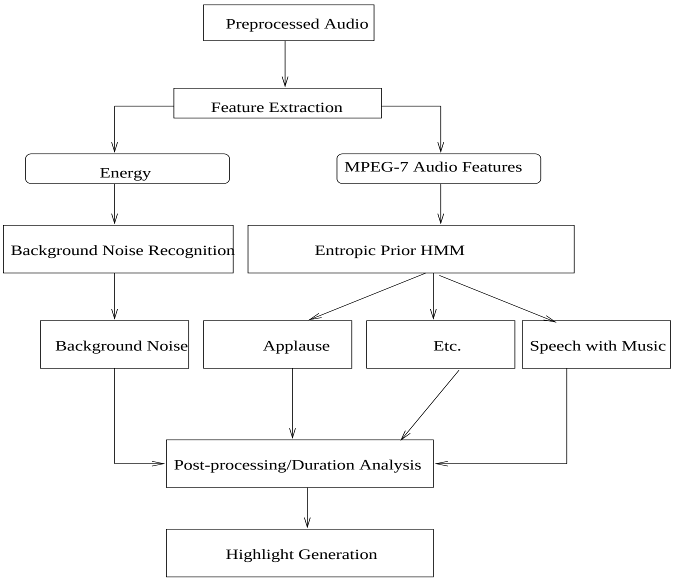

Fig. 1. Algorithm Flowchart

EP-HMM have been shown to improve the classification accuracy over ML-HMM on melody, text[7] and general sound classification[6].

4. PROPOSED APPROACH

Our approach mainly consists of three modules: (1) background noise recognition based on energy and magnitude, (2)audio classification using MPEG-7 audio features and Entropic Prior HMM and (3) post-processing for final presentation. An algorithmic overview is shown in Figure 1.

4.1. Background Noise Recognition

A silence detection algorithm is not appropriate in our framework as only golf games have background noise at low volume, but not baseball or soccer. Instead, we randomly pick $\frac{1}{100}$ of all segments of duration 0.5 seconds in the game's sound track and use their average energy and average magnitude as threshold to declare background noise segment.(Notice segments of silence can also be detected using this scheme.)

4.2. Feature Extraction and Audio Classification

In our feature extraction, an audio signal is divided into overlapping frames of duration 30ms with 10ms overlapping for a pair of consecutive frames. Each frame is multiplied by a hamming-window function.

The lower and upper boundary of the frequency bands for MPEG-7 features are 62.5Hz and 8kHz that are over a spectrum of 7 octaves. Each subband spans a quarter of an octave so there are 28 subbands in between. Those frequencies that are below 62.5Hz are group into 1 extra subband. After normalization of the 29 log subband energy, a 30-element vector represents the frame. This vector is then projected onto the first 10 PCA basis vectors of every class.

The basic unit for classification is a segment of audio of 0.5 seconds with 0.125 seconds overlapping. It is classified as one of the 6 classes each of which is modelled using a HMM. The class labels for all the segments of each game are passed to the post-processing module.

4.3. Post-processing

Because of the classification error, some post-processing scheme is needed to clean the labels returned by the classifier. We make use of the following observation: applause, cheering usually are of long durations(e.g, spanning over several continuous segments). Our post-processing technique is as the following: first, group continuous segments that are classified as applause or cheering respectively; then, declare segments longer than a certain percentage of the longest applause segments or cheering segments as applause or cheering.(We let this percentage to be a parameter provided to the end user. For the purpose of evaluating our experimental results, we have set it to 33%.)

4.4. Final Presentation

Applause or cheering usually takes place after some interesting play, either a good golf club-ball impact, baseball hit or a goal in soccer. The correct classification and identification of these segments allows the extraction of highlights possible due to this strong correlation.

Based on when the beginning of applause or cheering is, we output a pair of time-stamps, one being for a certain number of frames of video before and the other for after this starting point(Once again we let them be chosen by the end user). These time-stamps are used to display the video using random-access capability of most the state-of-the-art video players.

5. EXPERIMENTAL RESULTS AND ANALYSIS

5.1. Training and Testing Data Set

The training data is the same as that has been used in [8], including 90% of the 814 audio clips collected from TV broadcasting of golf, baseball and soccer games. The duration differs from around 0.5 seconds(for ball hit) to more than 10 seconds(for music segments) and the total duration is approximately 1 hour and 12 minutes.

The test data is composed of the audio tracks of 4 games(2 golf, 1 baseball, 1 soccer) of total duration 8.4 hours. Their durations are 2 hours, 1.4 hours, 3 hours and 2 hours respectively. For the two golf games, the background noise level of the first is low but high for the second because it took place on a raining day, the sound of the raining has been mixed into the audio track. The soccer game has high background noise from the excited soccer fans. The audio signals are all mono-channel, 16 bit per sample with a sampling rate of 16kHz.

	[A]	[B]	[C]	[D]	[E]	[F]	[G]
[1]	58	47	35	60.3%	74.5%	151	23.1%
[2]	42	94	24	57.1%	25.5%	512	4.7%
[3]	82	290	72	87.8%	24.8%	1392	5.2%
[4]	54	145	22	40.7%	15.1%	1393	1.6%

Table 2. Classification Results of the 4 games. [1]: golf game 1; [2]: golf game 2; [3] baseball game; [4] soccer game. [A]: Number of Applause and Cheering Portions(NACP) in Ground Truth Set; [B]: NACP by Classifiers WITH Post-processing; [C]: Number of TRUE ACP by Classifiers; [D]: Precision $\frac{[C]}{[A]}$; [E]: Recall $\frac{[C]}{[B]}$ WITH Post-processing; [F]: NACP by Classifiers WITHOUT Post-processing; [G]: Recall $\frac{[C]}{[F]}$ WITHOUT Post-processing.

5.2. Results

Deciding whether a highlight of a golf or soccer or baseball game is truly a highlight requires subjective assessment and is hence difficult. We therefore choose to use the classification accuracy of the applause and cheering since it is an objective criterion and is certainly strongly correlated with the presence of highlights. A high classification accuracy of these events leads to good highlight extraction. The applause or cheering portions of the 4 games are hand-labelled. Pairs of the beginning time and ending time stamps of these events are identified. They are the ground truth for us to compare with the classification results. The identification process is very time-consuming. It takes more than 4 hours of uninterrupted effort to establish the ground truth for an hour of audio.

Those 0.5 second-long segments that are continuously classified as applause or cheering respectively are grouped into portions. These portions are then checked to see whether they are true applause or cheering segments. The results are summarized into Table 2.

In Table 2, we have used ”precision-recall” to evaluate the performance. Precision is the percentage of events(applause or cheering) that are correctly classified. Recall is the percentage of classified events that are indeed correctly classified.

5.3. Comparison with Other Systems

In [1], 66 baseball highlights segments are selected by a human expert as ground truth; 49 highlights segments are identified by their algorithm, yielding a 75% precision. Our system yields a classification accuracy of $\frac{72}{82} = 87.8\%$. When we carefully choose the number of frames before the beginning of these correctly classified events to include the highlights, we can achieve this precision rate, or slightly lower.

In [2], 14 of the 21 test golf ball hit audio clips are correctly recognized by their algorithm, yielding a 66.7% precision. However, the audio signals have been carefully clipped from the game thus the boundary problem of the events is solved manually. Our recognition rate of golf ball hit events is lower, but we have dealt with the boundary problem directly.

For soccer, few audio-based highlights extraction systems have been reported in literature. The recognition rate of our system on soccer is worse than that on golf or baseball because it is difficult to distinguish cheering sound from the audience noise. The small difference between these two sounds is that there is a short surge of audio intensity at the beginning of cheering while the intensity is constantly high for audience noise. We notice that the missing of some of the cheering in soccer is quite acceptable such as those when the soccer ball changes the side of possession because these events are not so interesting as the goal shots. Our visual experience is very good when we watch the highlights generated even at this relatively low precision.

5.4. Discussion

We have shown that our approach for multiple sports highlights extraction is not less or just slightly worse than other approaches that target a specific sport. This makes it a serious candidate in sports highlights extraction applications with computational power constraints.

However, our precision rates are still low compared with the 94.728% recognition rate in Table 1. There are two possible reasons. At first, like the results in [2], this high recognition rate is achieved using testing clips that start with the onset of the event and end with the offset. When tested on audio signals where the boundaries of the highlights are not known, it is expected that the recognition rate will drop. Secondly, the existence of audio mixtures and noisy background degrades the performance of the event models trained using limited number of examples.

Post-processing techniques reduce false alarms as shown in the "[B]" column of Table 2. However, overall the false alarms rates are still high. There are also two possible reasons. Firstly, the trained models do not give 100% recognition accuracy so some of the targeted audio events will be mistaken for others. The other possible reason is that some of the audio signals such as bird chirping, raining or whistling can be misclassified as applause or cheering. We will address these issues in our future work.

6. CONCLUSION AND FUTURE WORK

We used MPEG-7 audio features and EP-HMM as the audio features and classifier to recognize these common audio phenomena for highlight extraction. Together with pre- and post-processing techniques based on some general sport knowledge, we have been able to generate promising results. We are exploring more audio features and better classifiers to increase the classification precision, decrease false alarms. In the future, we will also use video domain analysis techniques, such as detection of the score-change that is semantically related to highlights.

7. REFERENCES

[1] Y. Rui, A. Gupta, and A. Acero, "Automatically extracting highlights for TV baseball programs," *Eighth ACM International Conference on Multimedia*, pp. 105–115, 2000.

[2] W. Hsu, "Speech audio project report," *Class Project Report*, 2000, www.ee.columbia.edu/∼winston.

[3] L. Xie, S.F. Chang, A. Divakaran, and H. Sun, "Structure analysis of soccer video with hidden markov models," *Proc. Interational Conference on Acoustic, Speech and Signal Processing, (ICASSP-2002)*, May 2002, Orlando, FL, USA.

[4] P. Xu, L. Xie, S.F. Chang, A. Divakaran, A. Vetro, and H. Sun, "Algorithms and system for segmentation and structure analysis in soccer video," *Proceedings of IEEE Conference on Multimedia and Expo*, pp. 928–931, 2001.

[5] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[6] M. Casey, "Reduced-rank spectra and entropic priors as consistent and reliable cues for general sound recognition," *Proceeding of the Workshop on Consistent & Reliable Acoustic Cues for Sound Analysis*, 2001.

[7] M. Brand, "Structure discovery in conditional probability models via an entropic prior and parameter extinction," *Neural Computation*, vol. 11, no. 5, pp. 1155–1183, 1999.

[8] Z. Xiong, R. Radhakrishnan, A. Divakaran, and T.S. Huang, "Comparing MFCC and MPEG-7 audio features for feature extraction, ML-HMM and entropic prior HMM for sports audio classification," *submitted to ICASSP 2003*, April 6-10, 2003.