Improving Event detection in Cricket Videos Using Audio Feature Analysis

S. C. Premaratne [1], A. Gamanayake [1], K. L. Jayaratne [2], P. Sellappan[3]

[1] Faculty of Information Technology, University of Moratuwa, Kadubedda, Moratuwa, Sri Lanka
[2] S University of Colombo Schools of Computing, Philip Gunawardena Mawatha, Colombo 07, Sri Lanka
[3] School of Science and Technology, Malaysia University of Science and Technology (MUST), Malaysia
samindap@uom.lk, amila676@gmail.com, klj@ucsc.cmb.ac.lk, sell@must.edu.my

Received: October 30, 2020. Revised: April 5, 2021. Accepted: April 29, 2021. Published: May 13, 2021.

***Abstract* - This paper discusses an event detection approach based on audio, which proves to be effective when applied to the audio component of cricket television broadcaster's video. In this approach, both crowed noise levels which is in the background and commentator voice which is in focus are considered correlated to key events in the time frame. Classifiers related to Hidden Markov Model (or HMM) are utilized as it is an efficient tool for modeling processes varying with time and broadly used in the area of speech recognition. This experiment done using cricket television broadcaster videos, was successful thus this method can be used for automatic indexing of audio (or video containing audio) for quick searching or segmentation.**

***Keywords* - audio processing, voice detection, audio analysis, cricket key event detection, highlights.**

I. INTRODUCTION

TOPIC of sports/games video summarization or sport highlights detection is broadly used in researches, especially for sports like football and baseball. Previous studies have used audio [1-3] video[4,5], mixture of both audio-video[6], audio-textual indications, and a variety of features and classification techniques for summarization[6-8].

For a task of automatic broadcast summarization and highlights generation/finding using stored or live broadcasts, audios signals are the natural, top-most choice due to the light computational necessities and their ability in full filling the real-time requirements. In the same way that sports have different set of rules, playing conditions and atmosphere, the acoustic contents of each sport have set of distinguishable characteristics. The game we will discuss in this paper is Cricket, which can be characterized by players' chatting, shouting, applause, appeals, sound of ball hitting bat or timber etc. A game of basketball might have sounds of ball dribbling, referee whistles or sounds of ball bouncing from the board. Spectator clatter, Cheers made my cheer leaders and expert commentary are mutual content that can be found in most live sports broadcasts. Using or probing identified audio features, a researcher can automatically spot/generate summarizations or highlights for a broadcast video.

There are earlier studies and researches that took the approach of features, to detect highlights in multiple sports, unfortunately they were somewhat ineffective due to different sound patterns in crowd behavior, play noises, referee signals and even in commentary in different sports [9-11]. Even through cricket enjoy an extreme popularity especially in the Indian sub-continent, researches based on audio to detect highlights are rare. In this research, experiment was done on audio content of Cricket and a training mechanism will be developed and based on that highlight generation will be done. This approach can be used on both online and stored cricket-based sports content.

When compared with approaches of highlights generation in other sports, Cricket highlight generation using audio analysis will mainly present following challenges:

- In Cricket there is no referee whistle to indicate special events (Goals, Fouls, Penalties, etc.) of play.
- Crowed behaviors are significantly different in different parts of the world.
- As a single game of cricket might go through five days, even within the same match noise levels are different. Different commentary styles will be present with in same time frame.

To overcome the challenges, focus was mainly set reaching objectives

- Categorization of cricket match in to one of the 3 main formats (T20s) and weight audio analysis based on the cricket format.
- Audio events discovery with a negligible precision error and a high recall rate and.
- Generation of highlights/summery by using detected events.

A generic life cycle of Audio data analysis indicated below.

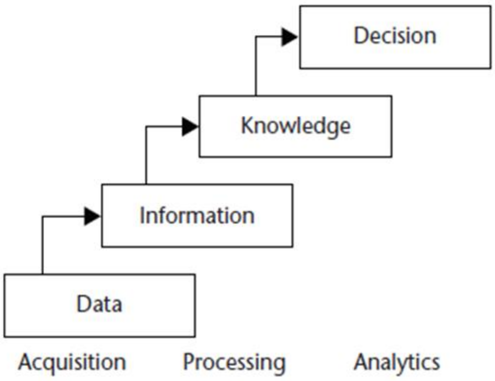

Fig. 1 Generic life cycle of audio data analysis.

Step 1: Data acquisition: - Acquiring data from different data sources.
Sept 2: Extract information: - Data preprocessing by extracting audio and other related features.
Sept 3: Generate knowledge: - usage of machine learning techniques to build knowledge models and evaluation process.
Sept 4: Decision: - Presentation of knowledge for decision making.

II. ACQUIRING AND PROCESSING

Cricket, historically being the gentlemen's game broadcasters always take the responsibility to create the ground atmosphere for the gentlemen who sit on their own sofa enjoying the game, audio has always been a very important feature of this presentations. Also, with newly found decision review (DRS) system accurate audio data has become more vital. Hence investment is done, and all required microphones are placed in the cricket pitch, on each of the stumps with the umpires sometimes even on players. Therefore, Audio data we find in television broadcasters audio stream is accurate and reliable for an experiment of this kind. Therefore, publicly available television broadcaster videos were processed using tool MATLAB to extract audio streams in "wav" format, at 44100 Hz, 16 bits per sample, and separated ball by ball. Using voiceActivityDetector, crowd noise and commentator speech were differentiated for each delivery.

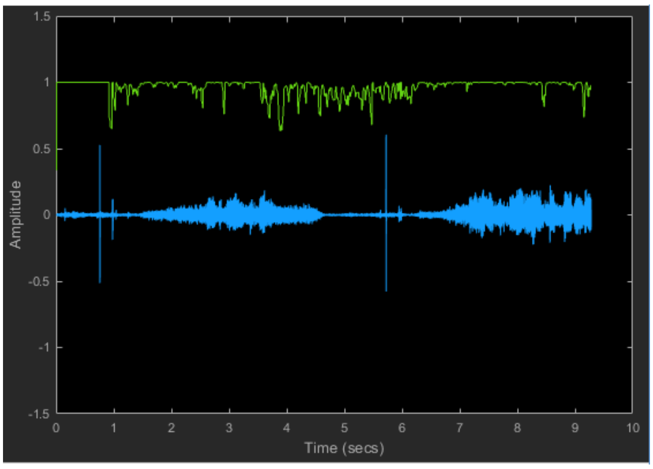

Fig. 2 Crowd noise and commentator speech variance for an example delivery

Where commentator speech is present, audio was further analyzed to determine whether it contains interesting words for highlight generations. Words such as boundary, four, six, maximum, gone, out, etc. This formed word set of words is evolving and improved as research moved on. This analyzing was done through MATLAB using IBM® Watson Speech to Text API and Audio Toolbox™ (Label Spoken Word functionality). A check is done to evaluate the presence of a known set of words in the audio and determine the time frame of the word is present. Each second of the video was labeled based on presence of the interesting word from commentary and that was used in pattern classes. Using this mechanism to generate a custom auto labeling function opened the door for a naval approach of combining interesting words in commentator speech with analyzed and obtained audio pattens.

Once the required data is extracted, classifiers based on Hidden Markov Model (HMM) were used to recognize acoustic patterns. Mel Frequency Cepstral Coefficients (MFCC) was used to parameterize above extracted Audio patterns.

III. PATTERN CLASSES

To classify the identified related pattern classes which are correspond to the possible key actions, a random sample of individual deliveries (span of one delivery – approximately 10 seconds) from random TV Broadcasts. Audio extract sampling was done at 44100Hz, in "wav" audio format was created. Then, the audio track of each delivery was split into separate observation sequences (length is one second). This training sample had 3600 such sequences, roughly around fifty minutes of audio (600 deliveries in cricket terms). To visualize the observations, the mean dimension was calculated for all features. Based on the findings following table was concluded.

I. Identified audio-based pattern classes

Class ID	Class Label	Class Description
0	N	No Commentary or Spectator Noise
1	CL	Interesting Commentary and Low/No Spectator Noise
2	L	Low Spectator Noise
3	CM	Interesting Commentary and Medium Spectator Noise
4	M	Medium Spectator Noise
5	CH	Interesting Commentary and High Spectator Noise
6	H	High Spectator Noise

From this experimental investigation, 7 demonstrative pattern classes were detected, even though class N No is not very useful in this experiment. Therefore, it will not be considered as an interested class and only the bottom 6 classes were considered from here onwards. Three interested classes were consisting speech and three of them does not. Again, out of interested six classes CL and L does not really represent an excitement in the game, one class consisting speech and the other class does not. In those time frames, there was minimal sound produced by the spectators present in the stadium. Classes ‘CM’ and ‘M’ represents time frames during a game that consists spectator noise such as music or chanting. While a match is going on it is not unusual (specially when organized groups of spectators are present.) for time frames of crowed chanting, usually these time frames are in between important events, such as a boundary or a wicket. It is vital to differentiate between spectator singing/chanting and crowds’ reactions correlated to any key moments of the game. Therefore, classes ‘CH’ and ‘H’, are used as a representation of spectator cheering. They are a combination of crowd noises, cheering, chatter, applause, and shouting, ideally triggered by a high-profile incident of the sport/game.

IV. MEL-FREQUENCY CEPSTRAL COFFICIENTS (MFCC)

Mel-Frequency Cepstral Coefficients (MFCC) is used to parameterize the audio track and extract required information. Mel-Frequency Cepstral Coefficients, broadly utilized in speech recognition[12]. Those are developed specifically to characterize speech. While being resistance to noise interruptions, it is ideal to differentiate between human speech and other sound components. Therefore, MFCC coefficients were an appropriate selection for the experiment, where the selected features are consisted.

V. HIDDEN MARKOV MODEL (HMM)

Hidden Markov Model (HMM) related classifiers the auditory pattern classes were modelled using a continuous density.[13,14] In this key problem is to decide is the selections of both the number of mixtures per state and optimal model size. Number of parameters to be estimated is dependent on number of states and mixtures per state. Therefore, if we are to reach effective classification accuracy, above discussed parameters should be projected with maximum possible accuracy. Trade off here is between improved model fit linked with larger and enriched models, but an exact and reliable parameter projection is restricted by the size of the selected training data.

As the count of selected parameters increases, amount of training samples needed for an accurate estimate also take an upward turn. An experiment to classify an appropriate amount of states and mixtures per state, with mixtures per state varying from 1 to 8 and states a variance from 1 to 15, when a pre-tagged training data set is used. 60% of the sample was utilized to train the models and 40% to generate the new predictive likelihood scores, where the predictive likelihood indicated how well a model "fits" to the data sample used. Still, due to limited training data, the covariance matrices were controlled to be transverse for each individual mixture. Using the acquired results, a 7 state Hidden Markov Model with 6 mixtures per state is selected as the optimal approach.

VI. DECISION PROCESS

It is more likely that key events were happened/occurred in the time frames that crowd is more vocal as those events are expected trigger crowed reactions. Those time frame will be categorized into previously defined classes ‘CH’ and ‘H’. Once a sequence of new observations has been classified, it is possible to identify possible highlights within the sequence. Every observation sequence is analyzed and classified in to one of the previously defined pattern classes based on highest model likelihood score. As this is an “Open world” problem a filtering process is put on place considering the scenarios that each model would not have been shown all possible outcomes. Further, by using a threshold possible outlier were identified. Observation sequences that were identified as outlier were placed in an ambiguous outlier class.

VII. EVALUATION

Occurrence of a key event usually causes a excitement among the spectators which lasts longer than couple of seconds and commentator speech consisting at least one word from predefined interesting word set, the length of an observation sequence, where a possible key event was identified if n sequential observations were classified as either ‘CH’ or ‘H’ (Feature Class ID 5 or 6). A visual representation of an identified key event, where the graph in figure 3 is 10 concurrent observation sequences grouped into one of the 7 (Classes IDs 0-7) categories is illustrated as figure 3. Graph illustrated in figure 4 indicates the location of a true event. This soundtrack enters the ‘CH’ class at 4 seconds and exits at 9 seconds. Which is a 4 seconds long possible key event. Therefore, this delivery contains a possible key event and it will be flagged as a Highlight.

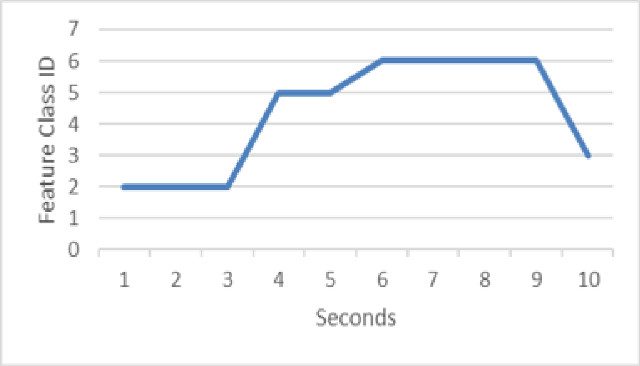

Fig. 3 Feature class variation for an example delivery

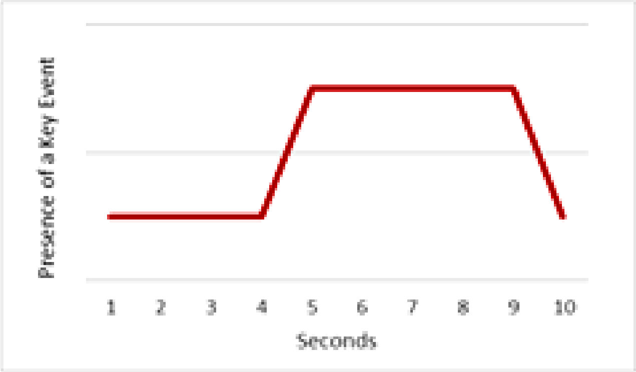

Fig. 4 Presence of a key event for an example delivery

Classifier was trained using a training set and then evaluated using a different set of deliveries. The second experimentation offers the results of the evaluation of the key event detection process fifty untrained deliveries. All these data were taken from international T20 game TV broadcaster videos which are publicly available.

A. Training and Test Data Sets

Classifiers were trained and evaluated using 600 separates, manually labelled, data samples (In cricketing terms deliveries). 60% of the collected data was used to train the classifiers and, 40% was used for the evaluation.

B. Event Detection

Ball by ball details were gathered from ESPN Cricinfo website (which is a publicly available) to evaluate the results. Important events such as Boundaries, Wickets, Appeals, Player milestones were considered as highlights that should be detected. Ball by ball reports indicated whether a specific delivery should be in highlight package or not. To measure performance, a properly identified event needs to be determined if a delivery that should be in highlight package is flagged/marked as a highlight.

II. Event Detection

	Actual	Detected	False
Highlight Deliveries	130	114	12
Other Deliveries	110	98	16

True Positives (TPs) - 114
False Negatives (FNs) - 16
False Positives (FPs) - 12
True Negatives (TNs) - 98

Following is the accuracy measures of the classifier based on data items used. Below calculations were made to evaluate the performance of the classifier through “F-Measure (F-Measure = (2 * Precision * Recall) / (Precision + Recall))” [15]

Precision = 114 / (114 + 12) = 0.9048
Recall = 114 / (114 + 16) = 0.8769
F-Measure = (2*0.9048*0.8769) / (0.9048+0.8769) = 0.8906

Comparing to the ball by ball information gathered by third party services, automatically detected/flagged highlight accuracy is high. But it is observed less important events that are in favor of the home team such a good ground fielding earning a considerable roar from the crowd and hence ending up as false detections. At the same time some key events which are in favor of the visiting team, such as player milestones were not cheered on and ended up as undetected.

VIII. CONCLUSIONS AND FUTURE WORK

Acoustic based highlight and key event detection discussed in the paper when applies to television broadcasters, main benefit is to generate a highlight package automatically. By applying Hidden Markov Model based classifiers to the data set, we were able to eradicate the need for defining a heuristic set of rules to identify key event detection thus avoiding a two-class approach, shown not to be appropriate. Hence, results obtained using the single HMM-based classifiers was inspiring given the difficult nature of the ball by ball audio and the limited amount of the training data, where the system overall detected 78% of the highlights from a new unseen collection.

In this experiment data was limited to T20 internationals to keep the nature of the audio tracks constant. But future plan is developing a system that can differentiate the kind of cricket (T20I, ODI or TEST) then classify accordingly. Also, in current system data need to be fed on ball by ball basis, therefore an important future improvement is to make the system work with the audio track of the entire match. On a concluding note, the event detection classification algorithms did fail to distinguish highlights coinciding with low crowd response. One probable resolution to this problem would be the addition of new features probably from different modalities such as video[16,17] or text[18].

References

[1] “Audio-Based Event Detection for Sports Video”, Mark Baillie, Joemon M. Jose, Department of Computing Science, University of Glasgow, 17 Lilybank Gardens, Glasgow, G12 8QQ, UK

[2] Baijal, A.; Jaeyoun Cho; Woojung Lee; Byeong-Seob Ko, "Sports highlights generation based on acoustic events detection: A rugby case study," IEEE International Conference on Consumer Electronics (ICCE), pp.20-23, 2015

[3] “An Efficient Ball Detection Framework for Cricket” B.L. Velammal[1] and P. Anandha Kumar[2]
1. Department of CSE, Anna University, Chennai, Tamilnadu 600025, India
2. Department of IT, MIT Campus, Anna University, Chennai, Tamilnadu 600025, India.

[4] “Detecting Highlights in Sports Videos: Cricket as a Test Case”, Hao Tang[1,2], Vivek Kwatra[2], Mehmet Emre Sargin[2], Ullas Gargi[2]
1. HP Labs, Palo Alto, CA USA
2. Google Inc., Mountain View, CA USA

[5] S. C. Premaratne, K. L. Jayaratne, P. Sellappan. A Novel Hybrid Adaptive Filter to Improve Video Keyframe Clustering to Support Event Resolution in Cricket Videos, International Journal of Engineering and Advanced Technology (IJEAT), 2019, ISSN 2249-8958.

[6] Qiang Huang; Cox, S., "Hierarchical language modeling for audio events detection in a sports game," IEEE ICASSP, pp.2286-2289, 2010

[7] Duxans, H.; Anguera, X.; Conejero, D., "Audio based soccer game summarization," IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, pp.1-6, 2009

[8] S. C. Premaratne, K. L. Jayaratne "Structural approach for event resolution in cricket videos" International Conference on Video and Image Processing (ICVIP) Singapore, Singapore - December 27 - 29, pp. 161--166 2017

[9] Z. Xiong, R. Radhakrishnan, A. Divakaran, and T.S. Huang, “Highlights extraction from sports video based on an audiovisual marker detection framework,” in IEEE ICME, Amsterdam, The Netherlands, June 2005, pp. 4–7.

[10] Zhou, W., S. Dao, and C.-C. Jay Kuo, On-line knowledge- and rule-based video classification system for video indexing and dissemination. Information Systems, 2002. 27(8): p. 559-586.

[11] Hasan, T., Bořil, H., Sangwan, A., and Hansen, J. H. "Multi-modal highlight generation for sports videos using an information-theoretic excitability measure". EURASIP Journal on Advances in Signal Processing, pp.1-17, 2013

[12] Tiwari, Vibha. (2010). MFCC and its applications in speaker recognition. Int. J. Emerg. Technol. 1.

[13] Weijian Li, Tongshun Liu, Time varying and condition adaptive hidden Markov model for tool wear state estimation and remaining useful life prediction in micro-milling, Mechanical Systems and Signal Processing, Volume 131, 2019, Pages 689-702, ISSN 0888-3270

[14] L. Rabiner and B. H. Juang. Fundamentals of Speech Recognition. Prentice Hall, Englewood Cliff s, NJ, USA, 1993

[15] Marina Sokolova, Guy Lapalme, "A systematic analysis of performance measures for classification tasks, Information Processing & Management", Volume 45, Issue 4, 2009, Pages 427-437, ISSN 0306-4573,

[16] M. R. Naphade, A. Garg, and T. S. Huang. Duration dependent input output markov models for audio-visual event detection. In ICME, Tokyo, Japan, August 2001. IEEE

[17] S. C. Premaratne, K. L. Jayaratne, P. Sellappan. Improving Event Resolution in Cricket Videos, ICGSP'18 Proceedings of the 2nd International Conference on Graphics and Signal Processing. Sydney, NSW, Australia. Pages 69-73. 2018

[18] V. Kobla, D. DeMenthon, and D. Doermann, "Identifying sports videos using replay, text, and camera motion features," in SPIE conference on Storage and Retrieval for Media Databases, 2000.