Audio-Based Event Detection for Sports Video

Mark Baillie and Joemon M. Jose

Department of Computing Science, University of Glasgow
17 Lilybank Gardens, Glasgow, G12 8QQ, UK
{bailliem,jj}@dcs.gla.ac.uk

Abstract. In this paper, we present an audio-based event detection approach shown to be effective when applied to the Sports broadcast data. The main benefit of this approach is the ability to recognise patterns that indicate high levels of crowd response which can be correlated to key events. By applying Hidden Markov Model-based classifiers, where the predefined content classes are parameterised using Mel-Frequency Cepstral Coefficients, we were able to eliminate the need for defining a heuristic set of rules to determine event detection, thus avoiding a two-class approach shown not to be suitable for this problem. Experimentation indicated that this is an effective method for classifying crowd response in Soccer matches, thus providing a basis for automatic indexing and summarisation.

1 Introduction

With the continual improvement of digital video compression standards and the availability of increasingly larger, more efficient storage space, new methods for accessing and searching digital media have become possible. A simple example would be the arrival of digital set top devices such as 'TiVo' [4] and 'Sky+' [16], that allow the consumer to record TV programmes straight to disk. Once stored, users can manually bookmark areas of interest within the video for future reference. Other advancements include Digital TV, where broadcasters have introduced interactive viewing options that present a wider choice of information to users. For example, viewers of Soccer can now choose between multiple camera angles, current game stats, email expert panelists and browse highlights, whilst watching a match. However, in order to generate real time highlights, it is necessary to log each key event as it happens, a largely manual process.

There has been a recent effort to automate the annotation of Sports broadcasts, which include the recognition of pitch markings [1, 3], player tracking [5], slow-motion replay detection [9, 17] and identification of commentator excitement [15]. Automatic indexing is not only beneficial for real time broadcast production but also advantageous to the consumer, who could automatically access indexed video once recorded to disk. However, current real-time production and in-depth off-line logging, required to index key events such as a goal, are on the whole manual techniques. It has been estimated that off-line logging, an in

E. M. Bakker et al. (Eds.): CIVR 2003, LNCS 2728, pp. 300–309, 2003.

depth annotation of every camera shot, can take a team of trained Librarians up to 10 hours to fully index one hour of video [8].

In this paper, we outline an approach to automatically index key events in Soccer broadcasts through the use of audio-based content classes. These content classes encapsulate the various levels of crowd response found during a match. The audio patterns associated with each class are then characterised through Mel-Frequency Cepstral Coefficients (MFCC) and modelled using Hidden Markov Model-based (HMM) classifiers, a technique shown to be effective when applied to the detection of explosions [11], TV genre classification [18] and speech recognition [14]. In Section 2, we will introduce the concept of event detection using audio information and, in Section 3 we evaluate the performance of our system, concluding our work in Section 4.

2 Audio-Based Indexing

Microphones are strategically placed at pitch level to recreate the stadium atmosphere for the armchair supporter[1]. As a result, the soundtrack of a Soccer broadcast is a mixture of speech and vocal crowd reactions, alongside other environmental sounds such as whistles, drums, clapping, etc. This atmosphere is then mixed with the commentary track to provide an enriched depiction of the action unfolding.

For event detection, we adopt a statistical approach to recognise audio-based patterns related to excited crowd reaction. For example, stadium supporters react to different stimuli during a match, such as a goal, an exciting passage of play or even a poor refereeing decision by cheering, shouting, singing, clapping or booing. Hence, an increase in crowd response is an important indicator for the occurrence of a key event, where the recognition of crowd reaction can be achieved through the use of Hidden Markov Model (HMM) based classifiers that identify audio patterns. These audio patterns are parameterised using Mel-Frequency Cepstral Coefficients (MFCC).

2.1 Feature Set

For this study, we selected Mel-Frequency Cepstral Coefficients (MFCC) to extract information and hence parameterise the soundtrack. MFCC coefficients, widely used in the field of speech detection and recognition (for an in-depth introduction refer to [14]), are specifically designed and proven to characterise speech. Also, MFCC have been shown to be robust to noise as well as being useful for discriminating between speech and other sound classes, such as music [2, 13]. Thus, as an initial starting point, MFCC coefficients were considered to be an appropriate selection for this problem, where the Feature Set consisted of 12 uncorrelated MFCC coefficients with the additional Log Energy [14]. Each

[1] An armchair supporter is a fan who prefers to view sport from the comfort of their armchair rather than actively attend the match.

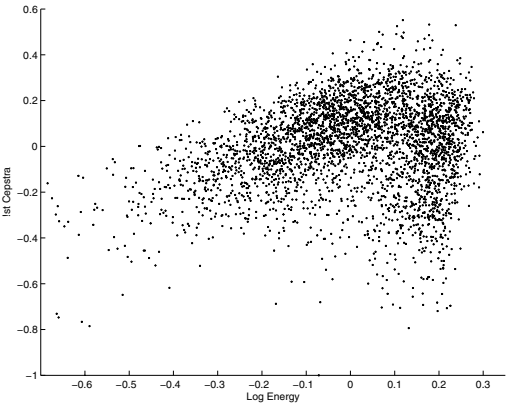

Fig. 1. Plot of the mean observation of Log Energy versus the 1st MFCC coefficient. There are two main clusters, the left containing observation sequences with speech, the right observations with no speech

Soccer broadcast was then split sequentially into one second observations, where the cepstra coefficients were computed every 10ms with a window size of 25ms, normalised to zero mean and unit variance.

2.2 Pattern Classes

An ideal solution to the problem of event detection would be a data set consisting of two content classes. One class made up of all audio clips that contained key events and the other class, the rest. But in reality this is not the case. Thus, in order to identify the relevant pattern classes that correspond to key events, we created a small random sample generated from 4 Soccer Broadcasts, digitally captured using a TV capture card. The audio track was sampled at 44100 Hz, using 16 bits per sample, in 'wav' format. Next, the soundtrack from each game was divided into individual observation sequences, one second in length. The training sample contained 3000 observation sequences, approximately 50 minutes of video. To visualise each observation, the mean measurement was calculated per feature.

Given the representative sample, scatter plots were created for all two dimensional Feature sub-space combinations, Fig. 1 is an example. From inspection of each plot, Fig. 1, it was clear that there were two main populations, those clips containing speech and those without, where each main group was a collection of smaller, more complex sub-classes. These sub-classes include differing levels of crowd sounds as well as the variation within and between the different speakers. Those clips containing high levels of crowd response, correlated to 'key events', were found to be grouped together, where in Fig. 1, these groups were positioned towards the 'top' of both main clusters. The data also contained a high frequency of outliers that through examination were discovered to be a mixture

Table 1. The selected audio-based pattern classes

Class Label	Class Description
S-l	Speech and Low Levels of Crowd Sound
N-l	Low Levels of Crowd Sound
S-m	Speech and Medium Levels of Crowd Sound
N-m	Medium Levels of Crowd Sound
S-h	Crowd Cheering and Speech
N-h	Crowd Cheering

of unusual sounds, not identifiable to any one group. These include signal interference, stadium announcements, music inside the stadium and also complete silence.

From this exploratory investigation, 6 representative pattern classes were selected, Table 1, where three of the classes contained speech and three did not. The first two classes, 'S-l' and 'N-l', represent a 'lull' during the match, one class containing speech and the other class not. During these periods, there was little or no sound produced from the stadium crowd. Classes 'S-m' and 'N-m', represent periods during a match that contain crowd sounds such as singing. During a match it is not unusual for periods of singing from supporters, usually these periods coincide with the start and end of the game, as well as after important events, such as a goal. Singing can also occur during lulls in the game where supporters may vocally encourage their team to improve performance. It is important for event detection to discriminate between crowd singing and those responses correlated to key moments during a game. Hence, the last two classes, 'S-h' and 'N-h', are a representation of crowd cheering. These classes are a mixture of crowd cheering, applause and shouting, normally triggered by a key incident during the game.

2.3 Hidden Markov Model-Based Classifiers

The audio-based pattern classes were modelled using a continuous density Hidden Markov Model (HMM). HMM is an effective tool for modelling time varying processes, widely used in the field of Speech Recognition (refer to [14], for an excellent tutorial on HMM). The basic structure of a HMM is: $\lambda = (A, B, \Pi)$, where A is the state transition matrix, B is the emission probability matrix and Π is the initial state probabilities. A HMM is a set of connected states $S = (s_1, s_2, \ldots s_n)$, where transition from one state to another is dependent only on the previous time point. These states are connected by transition probabilities $a_{ij} = p(s_i|s_j)$, where each state s_i has a probability density function, $b_{ij} = p(x|s_i)$ defining the probability of generating feature values at any given state. Finally, the initial state probabilities define the probability of commencing at any state given the observation sequence.

One difficulty when working with HMMs is model selection. For example, restrictions within A, the state transition probability matrix, can prevent move-

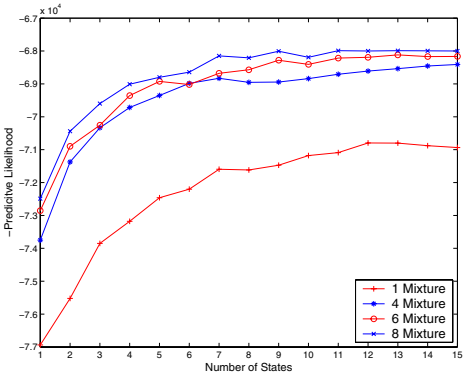

Fig. 2. Plot of predictive likelihood versus number of states

ment from one state to another, thus defining the behaviour of the model. A model that restricts movement from only left to right, is called a 'Bakis' Hidden Markov Model. This type of HMM can be very successful when applied to Automatic Speech Recognition [14], where each state(s) represents a phoneme in a word. Hence, as a sensible starting point, 'Bakis' HMMs were chosen to model each pattern class.

Another crucial issue to decide is the selection of both the optimal model size and number of (Gaussian) mixtures per state, where model size corresponds to the number of states. As the number of states and mixtures per state increase[1], so does the number of parameters to be estimated. To achieve successful classification these parameters must be estimated as accurately as possible. Note, there is a trade off in terms of better model fit associated with larger more enriched models, where precise and consistent parameter estimation is limited by the size and quality of the training data [6]. As the number of parameters increase so does the number of training samples required for accurate estimation.

To tackle this problem, we ran an experiment to identify a suitable number of states and mixtures per state. A number of 'Bakis' HMMs were generated, with states ranging from 1 to 15 and mixtures per state ranging from 1 to 8, using a pre-labelled training collection. 75% of the sample was used to train the models and 25% to generate the new predictive likelihood scores [7], where the predictive likelihood indicated how well a model 'fits' the data sample. The model parameters were then initialised, using the k-means segmentation process, and then re-estimated applying the 'Baum-Welch' Expectation-Maximisation algorithm. Note, because of the limited training data, the covariance matrices were constrained to be diagonal for each individual mixture. For each mixture number, Fig. 2, there was a 'levelling off' in performance at approximately 7 states.

[1] The number of computations associated with a HMM grow quadratically when increasing the number of states. That is $O(TN^2)$, where N is the number of states and T is the number of time steps in an observation sequence.

Increasing the number of mixtures per state, also produced a small improvement in performance. From the results, it was decided to select a 7 state HMM with 6 mixtures per state, since the increased number of parameters to be estimated using 8 mixtures outweighed the minimal improvement in model fit.

2.4 Decision Process

Once a sequence of new observations has been classified, we can then identify possible key events within the sequence, where a key event is likely to occur during periods of high crowd response, i.e. classes 'S-h' and 'N-h'.

Classification Given a new observation sequence, we measure the likelihood of it belonging to one of the 6 pattern classes, where the likelihood is determined using 'Viterbi's' decoding algorithm [14]. A new observation sequence is placed in the group that produces the highest model likelihood score. Given that this was an 'open world' problem and that each model would not have been shown all possible eventualities, a filtering process was required. The evidence of outliers during the exploratory analysis, Section 2.2, provided further proof of this requirement, so a threshold was introduced to flag possible outliers whose model likelihood scores exceed an experimentally set threshold. Flagged observation sequences were placed in a seventh 'ambiguous' outlier class.

Event Detection To identify key events, given a classified audio sequence, a further decision process was formed. Since a key event triggers a crowd response which normally lasts longer than 1 second, the length of an observation sequence, an 'event window' was introduced, where a key event was flagged if n sequential observations were classified as either 'S-h' or 'N-h'. Fig. 3 is an illustration of a detected event, where the top graph is 60 concurrent observation sequences grouped into one of the 7 categories. The bottom graph indicates the location of a true event. For this example we assume the 'event window' is set at 10 seconds. Hence, the soundtrack enters the 'S-h' class at 27 seconds and exits at 43 seconds, 16 observations later, thus flagging a key event. A correctly detected event was assumed to be an 'event window' that overlapped the location of a true event.

3 Experimental Results

This section outlines and presents the two experiments carried out to evaluate this approach to event detection. For the first experiment, each HMM-based classifier was first trained and then evaluated using a separate test set. The second experiment presents the results for the evaluation of the event detection process on two new unseen games.

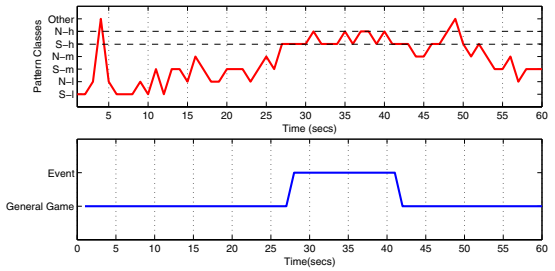

Fig. 3. Example of a detected event

Table 2. Labelled data for each class generated from 4 matches

Class	#1sec Observation Sequences
S-l	4020
S-m	1062
S-h	353
N-l	1832
N-m	545
N-h	213

Table 3. Confusion matrix for the HMM-based classification results

	Input - Actual					
Output Class	S-l	S-m	S-h	N-l	N-m	N-h
Outliers	1.21	4.43	7.12	5.02	4.3	10.76
S-l	**85.43**	13.32	1.11	3.21	0.01	0
S-m	9.5	**75.78**	4.3	0.2	2.94	1.2
S-h	3.5	4.21	**74.1**	0.65	0.24	4.21
N-l	0.33	1.23	0	**79.2**	5.32	5.34
N-M	0.03	0.4	1.22	9.6	**76.54**	9.81
N-H	0	0.63	12.15	2.12	10.65	**68.68**
Total	*100%*	*100%*	*100%*	*100%*	*100%*	*100%*

3.1 Classifier Performance

Each classifier was trained and evaluated on two separate, manually labelled, data samples, generated from 4 soccer broadcasts, see Table 2. Those content classes with high levels of crowd response contained lower numbers of samples in comparison with the other groups, due to the infrequency of key events. So, given these small numbers, 75% of the samples were used for training and 25% for testing and the performance of each classifier is presented in Table 3.1.

The two important classifiers for event detection, 'S-h' and 'N-h', representing high levels of crowd response, produced classification rates of 74% and 69%

Table 4. Event Detection

Class	#Key Events	#Detected	#False
Game1	24	20	8
Game2	16	14	7

respectively. For both classes, several of the observation sequences were misclassified as outliers. This may be an indication of the large intra-class variation within these two pattern classes. We also found an apparent overlap between the two groups 'N-m' and 'N-h', where a large number of observations from each group were falsely classified into the other class. This indicated a possible need for extending the framework into further, well defined sub-classes. Finally one common theme, indicated from the experiment, was that those models with a larger training sample performed better.

3.2 Event Detection Results

To measure the event detection approach, we gathered match reports and detailed game statistics for two new unseen games. The match reports were taken from 'OPTA' [12], a web-site dedicated to producing detailed summaries of soccer matches. Important events were considered to be goals, scoring attempts, cautions or other key incidents highlighted in the match report, forming the ground truth against which our system could be compared. The match reports also indicated approximate time points for each event, which aided this process. Using this information, a window from the start of the event to the end of the crowd response was created, for each true event.

To measure performance, a correctly identified event was determined to be: *"if a flagged 'event window' overlapped a 'true event window' at any time-point"*. If there was some overlap between an actual event and an 'event window', a correct detection was noted. If there was no true event during a flagged 'event window', a false detection was noted. For the experiment the 'event window' was experimentally set at 10 consecutive 1 second audio clips.

Comparing the automatically generated event index for the two games with the truth data from the match report, we found a high success rate, where only six events were not identified (Table 4). However, one of the missed events was a goal that was scored by the 'away' team, who were supported by a small section of the crowd in the stadium. The small support produced little crowd response in the stadium, thus the event was not detected by the system. Among the false detections were noticeable periods of singing from the stadium crowd. For example, after a 'goal', supporters sing for long periods, often triggering false events. Another interesting observation was one false event detection did in fact contain crowd cheering. During this period an amusing event occurred, triggering a large crowd reaction that was not reported in the match summary.

4 Conclusions and Future Work

The audio-based event detection approach outlined in this paper, was shown to be effective when applied to Soccer broadcasts, where the main benefit of the system was its ability to recognise patterns that indicate high levels of crowd response, correlated to key events. By applying HMM-based classifiers to the problem, we were able to eliminate the need for defining a heuristic set of rules to determine event detection thus avoiding a two-class approach, shown not to be suitable. Hence, the performance of the individual HMM-based classifiers was encouraging given the difficult nature of the Soccer soundtrack and the limited size of the training data, where the system overall detected 85% of the key events from a new unseen collection. Further experimentation is planned to train and test the system over a larger, more varied collection as well as compare the approach against other techniques.

The experiments also highlighted other potential improvements to this approach. These include the introduction of further representative classes that would manage the large variability found in a soundtrack, as well as further investigation into the development of model selection and the Feature set. For example, the test collection used in this study contained only male commentators, where a potential problem would be new broadcasts that contain female speech. Male and female voice is known to contain different characteristics, so future development will be required to identify new or modify current content classes to cope with various speakers from either gender.

In regards to model selection, the Bayesian Information Criterion (BIC) [10] is a technique that can be used to estimate the optimal model size, balancing predictive likelihood against model parameter size. Also, investigation and development into the identification of audio features, specifically suited for discriminating between the defined pattern classes, would be advantageous. Current research into audio-based content retrieval, differentiating between classes such as music and speech [2, 13], highlight this need.

On a final note, the event detection algorithm did fail to recognise key events coinciding with little to no crowd response. One possible solution to this problem would be the inclusion of new features possibly from different modalities such as vision or motion. Examples of classification using a combination of different modalities can be found in [11, 18], where a combination of visual and audio features was applied to the problems of explosion detection and video genre classification respectively.

Acknowledgements

Thanks to Prof. Keith van Rijsbergen, Prof. Mark Girolami, Robert Villa, Craig Hutchison, Marcos Theophylactou, Vassilis Plachouras, Sumitha Balasuriya and Tassos Tombros for their helpful advice, support and comments.

References

[1] Y. L. Chang, W. Zeng, I. Kamel, and R. Alonso. Integrated image and speech analysis for content-based video indexing. In *ICMCS*, pages 306–313. IEEE, 1996. 300

[2] D. Keislar E. Wold, T. Blum and J. Wheaton. Content-based classification, search, and retrieval of audio. In *In IEEE Multimedia*, volume 3, pages 27–36. IEEE, 1996. 301, 308

[3] Y. Gong, T. S. Lim, and H. C. Chua. Automatic parsing of tv soccer programs. In *ICMCS*, pages 167–174, Washington DC, May 1995. 300

[4] TiVo Inc. http://www.tivo.com/. Last visited 24th April 2003. 300

[5] S. Intille and A. Bobick. Visual tracking using closed worlds. Technical report, MIT Media Laboratory, 1995. http://web.media.mit.edu/ intille/. 300

[6] A. K. Jain, R. P. W. Duin, and J. Mao. Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37, January 2000. 304

[7] J. P. Cambell Jnr. Speaker recognition: A tutorial. In *Proceedings of the IEEE*, volume 85, pages 1437–1462, September 1997. 304

[8] J. Kittler, K. Messer, W. Christmas, B Levienaise-Obadia, and D. Koubaroulis. Generation of semantic cues for sports video annotation. In *ICIP*, pages 26–29, Thessaloniki, Greece, October 2001. 301

[9] K. Kobla, D. Doermann, and D. DeMenthon. Identification of sports videos using replay, text, and camera motion features. In *Conference on Storage and Retrieval for Media Databases*, volume 3972, pages 332–343. SPIE, January 2000. 300

[10] C. Li and G. Biswas. A bayesian approach to temporal data clustering using hidden markov models. In *ICML*, pages 543–550, Stanford, California, 2000. 308

[11] 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. 301, 308

[12] OPTA. http://www.opta.co.uk/. Last visited 24th April 2003. 307

[13] D. Pye. Content-based methods for the management of digital music. In *ICASSP*, volume IV, pages 2437–2400, 2000. 301, 308

[14] L. Rabiner and B. H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, Englewood Cliffs, NJ, USA, 1993. 301, 303, 304, 305

[15] Y. Rui, A. Gupta, and A. Acero. Automatically extracting highlights for tv baseball programs. In *ACM Multimedia*, pages 105–115, LA, 2000. 300

[16] Sky+. http://www.sky.com/. Last visited 24th April 2003. 300

[17] P. van Beek, H. Pan, and M. I. Sezan. Detection of slow-motion replay segments in sports video for highlights generation. In *ICASSP*, Utah, May 7-11 2001. 300

[18] Y. Wang, Z. Liu, and J. Huang. Multimedia content analysis using both audio and visual clues. In *IEEE Signal Processing Magazine*, volume 17, pages 12–36. 2000. 301, 308