Research & Development

White Paper

WHP 200

July 2011

Interestingness Detection in Sports Audio Broadcasts

Sam Davies

BRITISH BROADCASTING CORPORATION

© BBC 2012. All rights reserved.

White Paper WHP 200

Interestingness Detection in Sports Audio Broadcasts

Sam Davies

Abstract

This paper presents a novel method for semantic understanding of sports matches by extracting and ranking events within a match by interestingness. Using audio feature extraction, a system is presented which is able to segment between studio and pitch side broadcast. Key events within Rugby Union matches are then identified based on crowd excitation levels and referee whistles. This identifies individual interesting events and a timeline of interestingness estimation allowing viewers to navigate to sections of the broadcast where interesting sections of play occur.

This document was originally published in IEEE Proceedings of the Ninth International Conference on Machine Learning and Applications (ICMLA), Washington , USA. December 10-12 2010.

Additional key words: Accuracy, Estimation, Event detection, Feature extraction, Frequency estimation, Media, Noise

© BBC 2012. All rights reserved.

White Papers are distributed freely on request.

Authorisation of the Chief Scientist or General Manager is required for publication.

© BBC 2012. All rights reserved. Except as provided below, no part of this document may be reproduced in any material form (including photocopying or storing it in any medium by electronic means) without the prior written permission of BBC except in accordance with the provisions of the (UK) Copyright, Designs and Patents Act 1988.

The BBC grants permission to individuals and organisations to make copies of the entire document (including this copyright notice) for their own internal use. No copies of this document may be published, distributed or made available to third parties whether by paper, electronic or other means without the BBC's prior written permission. Where necessary, third parties should be directed to the relevant page on BBC's website at http://www.bbc.co.uk/rd/pubs/whp for a copy of this document.

White Paper WHP 200

Interestingness Detection in Sports Audio Broadcasts

Sam Davies

1 Introduction

The amount of digital audiovisual material available to viewers is growing rapidly as transmission capabilities increase and digital archived content is made accessible. With this, it is important that users are able to find the media and segments of media they want or could find interesting. To allow this metadata is required for each media asset allowing for inter and intra media asset searching.

Within the British Broadcasting Corporation (BBC), media assets that are likely to be reused have manually created metadata, contents of which range from brief synopses to detailed shot listings. However, this is a resource expensive process; a detailed analysis of a 30 minute programme can take a professional archivist 9 hours.

This manually generated metadata is for professional reuse. Frame accurate, it is designed to allow producers and researchers to find stock shots such as landscapes, key interviews or other clips. As the BBC open up their archives this level of detail and accuracy is not required by non-professional users; viewers. BBC Information and Archives (BBC I&A), the section of the BBC that archive programmes and create metadata, periodically release digitised collections of programmes online to the UK viewing public [1]. These collections are grouped by theme and have semantic metadata such as programme title, original transmission date, contributors and a brief synopsis.

Within these collections, media assets are already categorised and segmented manually; only those parts of programmes that are of relevance and interest are provided. As such, the synopses are all that is required for navigation and selection by viewers. This is a very labour intensive process and requires manual searching, tagging and segmenting of the archives. As more content is digitised this process will become less feasible. Thus automation will be required. Currently only a small fraction of the archives are digitised but internal digitisation projects such as [2] are continuing.

Digitised archive content is not the only problem space within this area. With the increased broadcast capacity offered by transmission channels such as digital and Internet Protocol (IP) based broadcasts, more content can be made available to the viewer at any time. Various UK broadcasters now offer ‘catch-up’ services, such as the BBC iPlayer [3], and there is work to integrate IP services with traditional set top boxes [4]. This ability to download and view vast amounts of content presents an increased requirement for search and selection. As viewers are presented with many programmes and have the ability to skip through, identifying which parts of a programme are interesting would be of great benefit. This identification would be of even greater use in large scale events, such as World Cups or the Olympics, where large numbers of competitions are broadcast concurrently or close together. These event detection tools could also be of use in production environments, providing a candidate list of events for inclusion in highlight shows.

This paper presents a novel method for segmenting sports broadcasts and for finding interesting events and sections of play within them from the audio. An interesting event is a single event that may be of interest. An interesting section is a passage of play that a viewer may want to watch. Continuous and discrete markers for interestingness are produced for the programme allowing viewers to navigate to specific events or general sections of interesting play. By using only

the audio track the system is able to distinguish between pitch side and studio commentary and then identify interesting events in pitch side segments. Further benefits of this audio only analysis is that radio broadcasts can also be segmented and without visual analysis, computational complexity is reduced.

Segmentation between pitch side and studio analysis is a key feature of this system. Sport is not solely shown within match environments. Many broadcasters present highlight programmes, which intersperse match segments with studio analysis. Viewers may not want to watch the studio analysis in these programmes instead just the match segments or vice versa. This system could offer the ability to segment between these, and then identify the key events within the match segments.

A heuristic based statistical analysis of the audio is used for segmentation and interestingness estimation, saving the computational expense of using other classification systems such as neural networks. The system could work on any sporting event where loud crowds react to play, without additional training. This paper presents an initial framework approach initially studying Rugby Union which could be extended for use with other sports such as football, American football or cricket.

This paper is organised as follows. Section II presents an overview of related research in audio feature extraction and event detection in sports. Sections III and IV present methods used for audio feature extraction and then segmentation and event detection respectively. Section V presents the preliminary results from this work, with these and future work discussed in section VI.

2 Background

Previous work in this area covers both sports and non sports event detection. These cover the selection of spectral and temporal features within the audio, and also methods for matching these low level audio features to higher level ones. Many different features have been studied and proposed for event classification in non-sports domains. Early research [5], looked at basic features such as loudness, pitch, brightness, bandwidth and harmonicity. Much interest has also centred around use of Mel Frequency Cepstrum Coefficients (MFCC) [6]. References [7, 8] both solely used MFCCs for feature extraction with [9, 10] combining these with other features such as those in [5] and spectral roll off, energy band ratio, Zero Crossing Rate (ZCR) and Short Time Frequency Estimation (STFE). Other interesting research has focussed on other psycho acoustical features. In a quite comprehensive analysis of features, [11, 12] compared a large set of standard low level features proposed by [10] with MFCCs and other psychoacoustic features such as loudness, roughness and sharpness finding their Auditory Filterbank Temporal Envelopes (AFTE), based on gammatone filters were the most effective. Reference [13] took a similar filter bank approach, using the Bark scale. These systems all aimed to create generic audio event classifiers. Features identified in these works are used in the initial study of classifiers, described in section III.

Other research has focussed on sport event detections. Reference [14] proposed an event detector for football, linking stadium noise and commentator speech to key events using MFCCs and log spectral energy. This was an extension to previous work [15] which only identified crowd cheering. In these works they related excited crowd noise to exciting events in a match which this paper builds upon. Commentator speech is thought to be of less importance as in much BBC sport analysed, commentators remained fairly neutral. MFCCs were also the basis of [16], which combined them with Perceptual Linear Prediction initially discriminating between speech and non-speech then classifying accordingly. Reference [17] used Linear Spectral Pairs (LSPs) and STFE, in Eurosport broadcasts, showing effective speech\non-speech discrimination. However, speech continues throughout broadcasts studied and so this discrimination is of less interest. Extending the scope of features used [18] looked at using MFCCs, ZCR, STFEs, sub-band energy, bandwidth and pitch frequency. Whilst this paper just looked to segment sport into play, adverts and studio segments, they found that usage of ZCR was important in speech/music discrimination, key for differentiating between adverts and programme. This is not an issue here as the BBC do not show

© BBC 2012. All rights reserved.

adverts. In terms of identifying actual events and features, [19] used a combination of MFCCS, Linear Predication Coding Coefficients (LPCCs), ZCR and STFEs. These were used to generate 'audio keywords', an attempt to match the low level features to general and specific semantic events in several sports such as goals. Identification of key events is an interesting concept, yet the system described in this paper any generated interesting event, not predefined ones. Looking for specific events is also a theme explored by [20]. Here, along with commentator speech excitation levels, they looked at the impulsive noise of a baseball bat hitting the ball. Whilst this would give a strong indicator as to when a strike was made, it limits the system to a set category of sports and to those in which there are distinct noises.

Use of audio feature extraction is not the only method for event detection. Reference [21] used MPEG 7 audio descriptors for the modelling of several sports. Whilst this offers a robust solution, the vast majority of broadcasts do not use the MPEG 7 framework. Reference [22] took an altogether different approach by looking at closed captions on screen. This was shown to be effective, but required the presence of captions which are not universally available in archived material. Reference [23] took a unified approach, using speech recognition to identify key words in the commentary and then match these with crowd excitation levels. Whilst shown to be effective, accurate and detailed training was required. Also, in many matches studied the commentators do not always explicitly name the event. Alternative approaches are to take a crowd based approach. Reference [24] provides an interesting approach based upon synchronisation of user generated media. Reference [25] also takes a similar approach; identifying which sections of media are interesting by identifying how many matching clips are uploaded onto social media sites. However, the system presented here is designed to allow for archives to be opened up. As such the material to be analysed often isn't available on any social media site, and so these approaches would not be appropriate.

3. Audio Feature Extraction

From an analysis by the authors of features taken from [5, 6, 9, 10, 12, 13, 26], three are identified as being most suitable; short time fundamental frequency estimation, signal power spectral density, and a 20 bank gammatone filter, summing banks one to four and separately bank six.

a. Short Time Fundamental Frequency Estimation

Short time fundamental frequency estimation is based upon the autocorrelation of the incoming signal. The 500ms windowed signal is further sub windowed into non overlapping 2048 sample length sub windows (*w*), giving a discrete time signal *x(n)*. The fundamental frequency f_0 is then taken as the reciprocal of the maximum value from the autocorrelation function *r(τ),* where *τ* is the time lag. The 2048 sub windows are then combined and averaged to give an estimation for the entire window (1)

$$r(\tau) = \frac{1}{N}\sum_{n=0}^{N-n-1} x(n)x(n+\tau)$$

$$f_0 = \frac{1}{W}\sum_{w=1}^{W}\left(\frac{1}{\max(r(\tau)_w)}\right) \quad (1)$$

b. Power Spectral Density

This is a measure of amplitude for different frequencies. For this, Welch's Method [27] is used for estimating the audio signals power as a function of frequency. This gives a spectrum of power for frequencies, from which the mean is calculated. This is used to give an indication as to the signal's power as a function of frequency.

© BBC 2012. All rights reserved.

c. First to Fourth and Sixth Gammatone Filters

Based on [11, 12] use of AFTEs discussed in section II, the incoming signal is put through a bank of 20 4th order Gammatone filters [28, 29]. Like MFCCs and the Bark Scale [30], this splits the incoming signal into overlapping filterbanks with logarithmically spaced central frequencies aligned with the auditory response of the human ear. A Hilbert transform is taken of the signal giving the temporal envelope. Principle component analysis (PCA)[31] is used to decompose this and the PCA values for the first four banks summed together to identify crowd noise and the sixth banks PCA value used for referee whistles.

4. Segmentation and Event Detection

This system needs no training to identify interesting events using instead a heuristic approach to a statistical analysis of the audio. The match is initially segmented between studio based (i.e. pre and post match analysis) and pitch side based sections (match play). Once these distinctions are made, any section identified as being pitch based is further analysed for interesting play. In order to segment the play, the autocovariance of the short time fundamental frequency estimations are averaged together over a 10 second sliding window offset by 1 second. An 8th order integrating polynomial is calculated as a line of best fit which maps the confidence values for studio based segments resulting in three peaks relating to pre match, half time and post match analysis (Fig. 1).

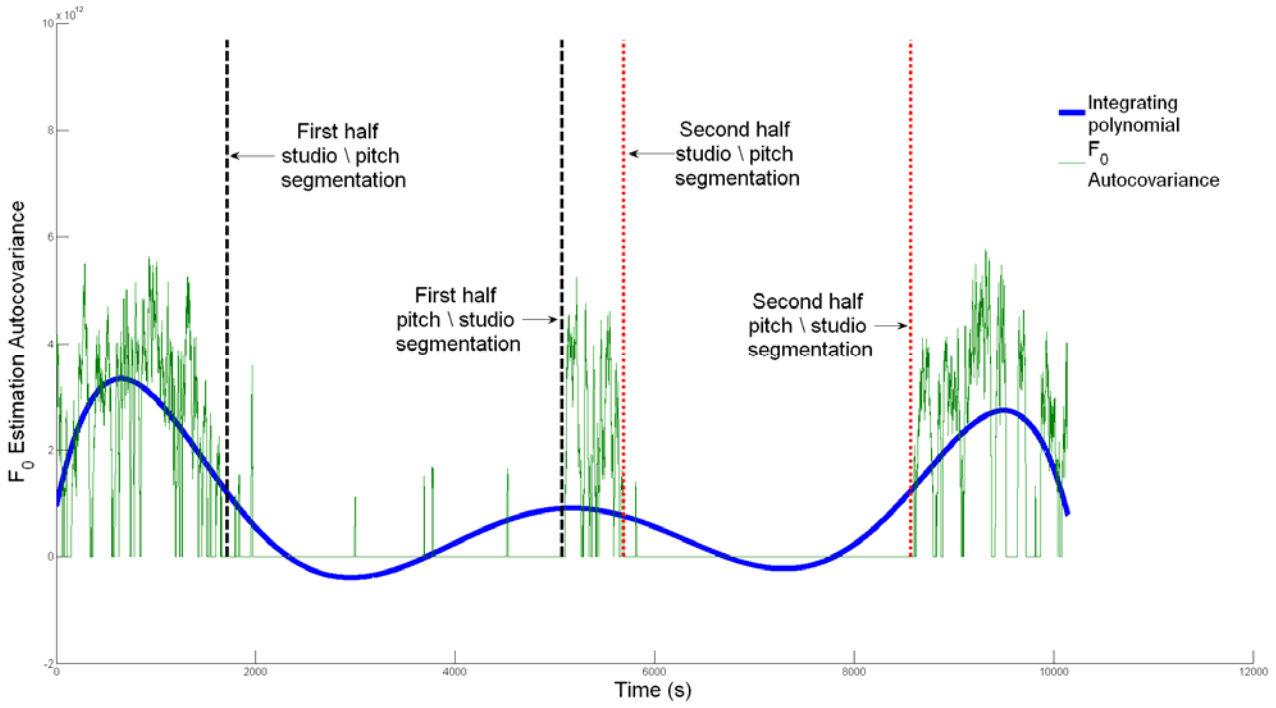

Figure 1. F_0 autocovariance with 8th order integrating polynomial showing segmentation

During studio sections there is usually only talking as experts analyse previous play. On average this gives a higher pitch estimation autocovariance, as gaps between words and speakers are noticeable, altering the estimated pitch and increasing variance. When this discussion moves to the pitch side, the constant background noise of the crowd and stadium fill in these gaps, reducing pitch variance dramatically. A high pass filter was also used on the estimation autocovariance to aid segmentation.

© BBC 2012. All rights reserved.

On the line of best fit, Half Maximum Amplitude (HMA) is used as an initial segmentation indicator. Once this is identified, the system then looks for crowd noise to help identify a more accurate segmentation point. Whilst the autocovariance of the short time frequency estimation is a robust identifier for segmentation, crowd noise can help identify interesting points near the start of each half such as teams appearing on the pitch or at the end such as final whistles. To find the start of each match, an inverse exponential weighting is applied to identified crowd noises finding the highest value crowd noise closest to the HMA indicator. To find the end of each half of the match exponentially weighted crowd noise events are used to find the greatest peak between the HMA value and the peak value (p_2) for the first half and for the second. The start of the second half is identified in a similar fashion as the start of the first half, identifying the event with maximum value after the interval peak value and the following HMA value.

Once the audio is segmented into pitch and studio segments the next stage is detecting unusual events, taking a two stage approach. The 6th AFTE PCA decomposition value readily shows referee whistles. As can be seen (fig. 2) when a whistle is detected on the audio, there is a clear peak in this band. This on its own is not sufficient for event detection. In many instances an event does not coincide with a referee whistle or is not audible so not identified. Thus crowd noise is also used. This is a much better indicator for event detection. Crowds at large sporting events always react to interesting events, be they scores, score opportunities or spectacular play. The length of the crowd cheers differs depending upon the situation, relating to the length and value of interestingness of the play. From an analysis of the events in matches identified by the authors, a cheer of one second was found to be the minimum length for an interesting event. AFTE bands one to four are found to map very closely to crowd excitation levels. As can be seen (fig. 2), there is a marked increase in these bands when an exciting event occurs. This is also found in the mean of the PSD amplitudes, averaged over 1 second.

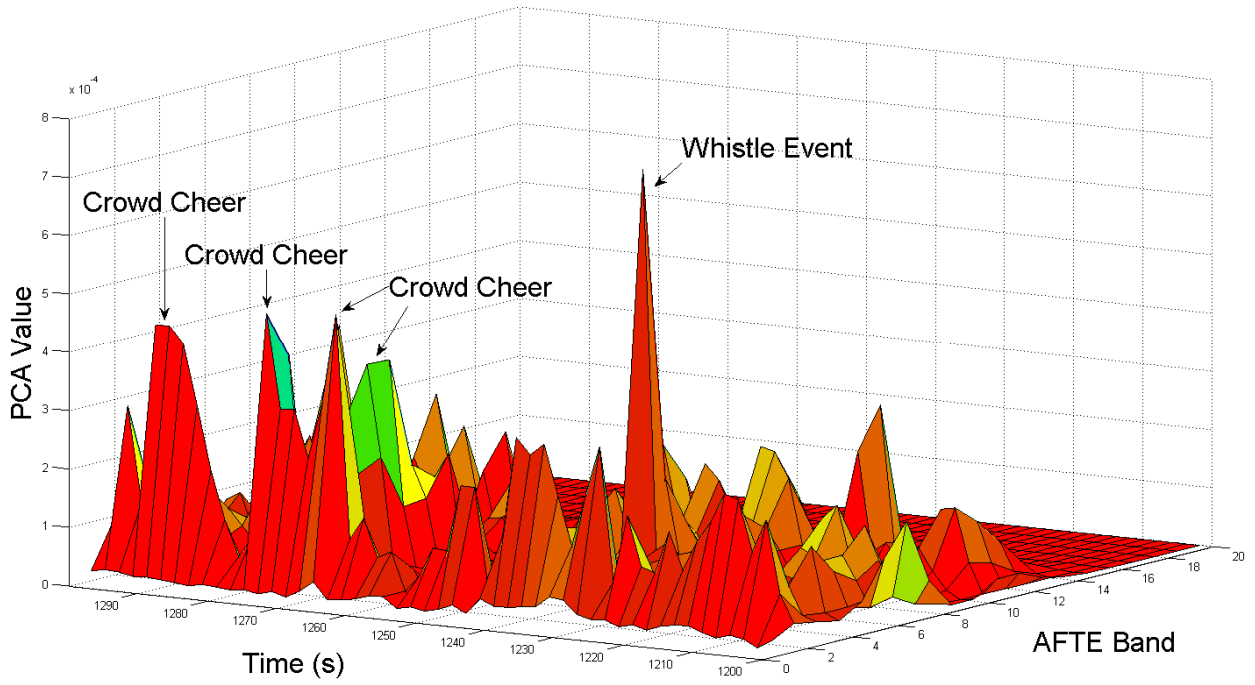

Figure 2. AFTE bands showing crowd excitation and referee whistle

An unusual event is declared if a summation of the lowest AFTE bands and PSD value, or lowest AFTE bands, PSD value and referees whistle rise above the mean value plus one standard deviation. This created a list of possible interesting events and respective confidence levels for the entire play segment.

© BBC 2012. All rights reserved.

As the maximum section of play that is analysed is 1 second if an interesting event occurred over several seconds many events are detected. These require filtering when identifying events. As the section of play becomes more interesting the confidence levels for the detected events rise. Conversely as the section of interesting play passes, confidence levels fall giving clear segments of interesting play. Final confidence levels are given as the integrals of these individual segments, and the time of the peak of these segments (or first peak if a plateau is found) used as the time of the interesting event.

Interestingness timelines are created before the filtering, giving continuous values for interestingness as opposed to discrete events. This gives a higher level view of the match, identifying segments of play as interesting. For example, in fig. 3 there are clear peaks in interestingness from around the 105th minute to 122nd. This shows a section of a match between Scotland and England in the 2010 RBS Six Nations Championship. This section relates to a penalty awarded to Scotland for foul play (crowd boos) before the penalty is missed (crowd groans). This is followed by an attack by Scotland (crowd cheers) before England defence stop the attack (crowd cheers). However, Scotland are awarded a second penalty (crowd cheers) which is scored (crowd cheers).

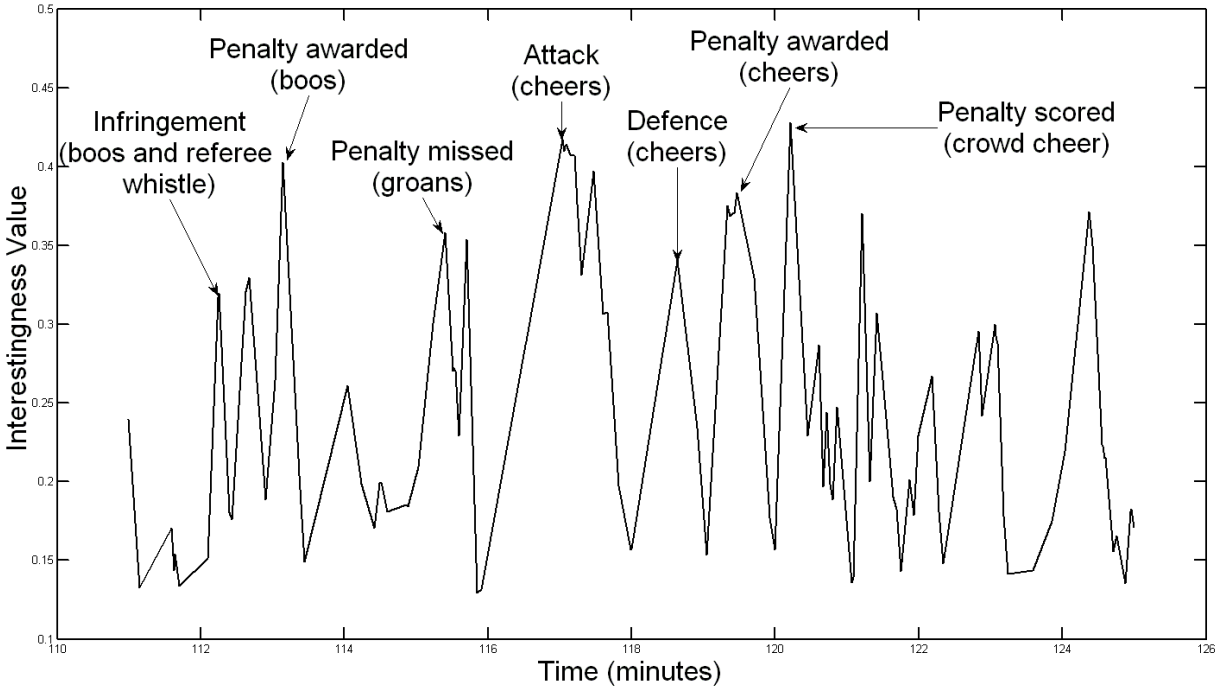

Figure 3. Interestingness timelines for rugby match

5. Results

Initial testing is to ascertain if the segmentation between pitch side and studio analysis performs well. The system is not looking to identify any specific event in this segmentation, just one that is before the start of play and after the switch from studio to pitch side. This is found to be the least successful part of the system and of the 48 different segments accuracy of 70.2% is achieved.

The next stage in testing is to ascertain if the system identifies exciting events. This is measured in two ways. The BBC Sports Library often provides detailed event logs for matches that have been broadcast. Whilst these are primarily aimed at production reuse (featuring non-play related information), they also contain interesting events, labelled ‘GOOD’ followed by an event description. Of the 24 matches studied, 6 have these logs. These are identified in table I.

© BBC 2012. All rights reserved.

TABLE I. RESULTS FOR LOGGED MATCHES

Match Name	System detected events	Professionally identified events	Accuracy
Ireland vs. France	21	33	63.63%
Scotland vs. Wales	25	30	83.33%
France vs. Scotland	29	30	96.66%
Italy vs. Ireland	23	37	62.16%
Wales vs. Ireland	21	22	95.45%
England vs. Scotland	26	33	78.78%

Of the 185 events identified by BBC Sport to be of interest in these matches, the system identifies 145 of them, giving a success rate of 78%. The remainder of the matches were checked against event logs made by the authors. The system is very successful at identifying interestingness, with 86% of interesting events identified. However, this high success rate also gave many false positives with 23% of all events false positives. Due to copyright reasons, it is not possible to release these professionally created event logs.

The final stage of testing is in the creation of interestingness timelines for each match as in fig. 3. These are found to be a very accurate indicator of where key segments of interest begin, peak and end. As can be seen from fig. 3, there are clear peaks in the graph, which correspond to interesting events such as tries and penalties. One of the more successful aspects of the graph is in the spike of interest event detection around these key events. These are clearly visible as groups of peaks, as shown in (fig. 3). In rugby and many sports, it is not just the scoring opportunity that is of interest, but in the play that builds up to this, which these timelines identify; distinct events are relatively uncommon in isolation.

6. Conclusions and Future Work

A system has been presented which segments and provides interestingness detection in sports audio. The system shows accurate segmentation between studio based discussions and pitch side segments. The system also shows accurate detection of interesting events.

There are instances where an interesting segment of play is manually identified that the system does not detect, as crowd reaction was either too brief or nonexistent. Without complete accuracy, production staff and archivists will still have to manually go through the data rendering this system less effective for this. However for opening up archives to viewer, the level of accuracy provided here is sufficient.

One of the key results from this work is in the identification of an interestingness timeline. This would give viewers a very clear estimation of where interesting sections, rather than events, occur; the more peaks there are, the more interesting that section. This is valuable tool for viewers to skip to interesting sections allowing the build up and the event, rather than just the event to be viewed.

Two areas for improvement in the system are in the improvement of segmentation accuracy and reduction of false positives. The majority of the false positives occur when the system incorrectly identifies a studio based segment as a pitch side segment. Here, replays not actual events are classified as interesting events. Improving the accuracy of the segmentation will reduce this. The system deliberately uses a simplistic approach to reduce computational complexity. Investigation into the use of more advanced machine learning methods could provide improved accuracy and is planned for future work.

© BBC 2012. All rights reserved.

Other further work is to also look at an implementation on archived content and how changes in production methods and audio quality affect the system. Modern sports productions consist of many different microphones around the stadium, which pick up crowd and referee whistles well which archived matches may not.

Other work is also planned to increase the number of sports the system can work with. One of the goals in not using any training data was to make the system as heterogeneous as possible detecting interestingness in any major sport where there is crowd and referee noise.

Video is not going to be considered as a further area of research. This system is designed to allow for analysis on audio only – which would be of far greater use in the BBC Archive where along with TV broadcasts of sport, there are many thousands of hours of radio broadcasts.

One key area of further research is whether the audiences want to navigate through sports content in this manner. This is a major area of work in which the BBC will shortly begin investigation.

7. Acknowledgements

This work was completed as part of the BBC Research and Development Multimedia Classification project [26].

8. References

[1] BBC. (2010, 17th July). *BBC Archive Collections*. Available: http://www.bbc.co.uk/archive/collections.shtml

[2] S. Cunningham and P. de Nier, "File-based Production: Making It Work In Practice," in *Proceedings of the International Broadcasting Convention*, Amsterdam, NL, 2007.

[3] BBC. (2010, 17th July). *BBC iPlayer*. Available: http://www.bbc.co.uk/iplayer/

[4] P. Canvas. (2010, 17th July). *Project Canvas*. Available: http://www.projectcanvas.info/

[5] E. Wold, T. Blum, D. Keislar, and J. Wheaton, "Content-Based Classification, Search, and Retrieval of Audio," *IEEE MultiMedia,* vol. 3, pp. 27-36, 1996.

[6] P. Mermelstein, "Distance measures for speech recognition, psychological and instrumental," *Pattern Recognition and Artifical Intelligence,* pp. 374-388, 1976.

[7] J.-J. Aucouturier, F. Pachet, P. Roy, and A. Beurivé, "Signal + Context = Better Classification.," presented at the Proceedings of ISMIR 07, Vienna, Austria, 2007.

[8] M. Cooper and J. Foote, "Automatic Music Summarization via Simliarity analysis," presented at the Proceedings of ISMIR 02, Paris, France, 2002.

[9] S. Bray and G. Tzanetakis, "Distributed audio feature extraction for music," presented at the Proceedings of ISMIR 05, London, UK, 2005.

[10] D. Li, I. K. Sethi, N. Dimitrova, and T. McGee, "Classification of general audio data for content-based retrieval," *Pattern Recognition Letters,* vol. 22, pp. 533-544, 2001.

[11] J. Breebaart and M. McKinney, "Features for Audio Classification," presented at the Proceedings of the Philips Symposium of Intelligent Algorithm, Eindhoven, NL, 2002.

[12] M. McKinney and J. Breebaart, "Features for audio and music classification.," in *Proceedings of ISMIR 03*, Baltimore, USA, 2003.

[13] E. Pampalk, "A Matlab Toolbox to Compute Music Similarity from Audio," presented at the Proceedings of ISMIR 04, Barcelona, Spain, 2004.

[14] M. Baillie and M. Jose, "An Audio-Based Sports Video Segmentation and Event Detection Algorithm," presented at the Proceedings of 2004 CVPR Washington, USA, 2004.

[15] M. Baillie and M. Jose, "Audio Based Event Detection for Sports Video," *Lecture Notes in Computer Science,* vol. 2728, pp. 61-65, 2003.

[16] J. Portelo*, et al.*, "Non-speech audio event detection," in *ICASSP 2009*, Taipei, Taiwan, 2009, pp. 1973-1976.

[17] H. Jun, D. Yuan, L. Jiqing, D. Chengyu, and W. Haila, "Sports audio segmentation and classification," in *IC-NIDC 2009*, Beijing, China, 2009, pp. 379-383.

© BBC 2012. All rights reserved.

[18] L. Bai, S. Lao, H. Liao, and J. Chen, "Audio Classification and Segmentation for Sports Video Structure using Support Vector Machines," in *2006 ICMLC* Dalian, China, 2006.
[19] M. Xu, C. Xu, L. Duan, J. S. Jin, and S. Luo, "Audio keywords generation for sports video analysis," *ACM Transactions on Multimedia Computing, Communications and Applications,* vol. 4, pp. 1-23, 2008.
[20] Y. Rui, A. Gupta, and A. Acero, "Automatically extracting highlights for TV Baseball programs," presented at the Proceedings of the eighth ACM international conference on Multimedia, California, United States, 2000.
[21] X. Ziyou, R. Radhakrishnan, A. Divakaran, and T. S. Huang, "Audio events detection based highlights extraction from baseball, golf and soccer games in a unified framework," in *Proceedings of ICME 2003*, Baltimore, USA, 2003, pp. III-401-4 vol.3.
[22] J. Cheolkon, L. Su Young, and K. Joongkyu, "Robust detection of key captions for sports video understanding," in *Proceedings of ICIP 2008*, California, USA, 2008, pp. 2520-2523.
[23] Y. Chang, W. Zeng, I. Kamel, and R. Alonso, "Integrated Image and Speech Analysis for Content-Based Video Indexing," presented at the Proceedings of the 1996 ICMCS, Hiroshima, Japan, 1996.
[24] L. Kennedy and M. Naaman, "Less talk, more rock: automated organization of community-contributed collections of concert videos," presented at the Proceedings of the 18th international conference on World wide web, Madrid, Spain, 2009.
[25] J. S. Pedro, V. Kalnikaite, and S. Whittaker, "You can play that again: exploring social redundancy to derive highlight regions in videos," presented at the Proceedings of the 13th international conference on Intelligent user interfaces, Sanibel Island, Florida, USA, 2009.
[26] D. Bland, S. Davies, and N. Pinks, "Generating Metadata from AV Content," UK Patent, 2010.
[27] P. Welch, "The Use of Fast Fourier Transform for tghe Estimation of Power Spectra: A Method Based on Time Averaging Over Short, Modified Periodgrams," *IEEE Transactions on Audio Electroacoustics,* vol. AU-15, pp. 70-73, 1967.
[28] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research,* vol. 47, pp. 103-138, 1990.
[29] R. Patterson*, et al.*, "Complex Sounds and Auditory Images," in *Proceedings of the 9th International Symposium on Hearing*, Oxford, UK, 1992, pp. 429-446.
[30] E. Zwicker, "Subdivision of the Audible Frequency Range into Critical Bands (Frequenzgruppen)," *The Journal of the Acoustical Society of America,* vol. 33, p. 248, 1961.
[31] L. Smith, "A tutorial on Principal Components Analysis," unpublished.

© BBC 2012. All rights reserved.