Automatic Excitement-Level Detection for Sports Highlights Generation

Hynek Bořil, Abhijeet Sangwan, Taufiq Hasan, John H. L. Hansen

Center for Robust Speech Systems (CRSS), Eric Jonsson School of Engineering,
University of Texas at Dallas, Richardson, Texas, U.S.A.
http://crss.utdallas.edu

Abstract

The problem of automatic excitement detection in baseball videos is considered and applied for highlight generation. This paper focuses on detecting exciting events in video using complementary information from the audio and video domains. First, a new measure for non-stationarity which is extremely effective in separating background from speech is proposed. This new feature is employed in an unsupervised GMM-based segmentation algorithm that identifies the sports commentators speech within the crowd background. Thereafter, the "level-of-excitement" is measured using features such as pitch, F_1–F_3 center frequencies, and spectral center of gravity extracted from the commentators speech. Our experiments using actual baseball videos show that these features are well correlated with human assessment of excitability. Furthermore, slow-motion replay and baseball pitching-scenes from the video are also detected to estimate scene end-points. Finally, audio/video information is fused to rank-order scenes by "excitability" in order to generate highlights of user-defined time-lengths. The techniques described in this paper are generic and applicable to a variety of topic and video/acoustic domains.

Index Terms: Video Segmentation, Multimodal Signal Processing

1. Introduction

This study focuses on the problem of identifying exciting-events in multimedia content. Our approach analyzes speech characteristics that identify islands (or "hot-spots") of strong emotion. In general, the ability to automatically parse multimedia content and tag "interesting events" is important for many domains such as sports, security, movies/TV shows, broadcast news, *etc.* A number of technologies such as search, summarization, and mash-ups, can utilize "hot-spot" information to enhance access to, as well as navigation of content. For example, emotional "hot-spots" within sports videos are very likely to be "exciting" and this information can be used to guide the process of automatically generating highlights. This constitutes the motivation for this work, where automatic highlights of baseball videos are generated using emotional "hot-spot" detection (or "exciting events" detection).

Researchers have utilized audio and video streams to extract features that identify exciting plays in sports videos. Among video-based features, motion and density of cuts have been found to be useful for detection [1]. On the other hand, audio-based features have been derived from both speech (generally commentators) and background (generally audience), where audience-events like cheering/applause as well as the commentators speech characteristics have proven to be useful [2, 3]. While video-based features tend to be more game-dependent, audio-based features (audience and commentators) are more generic and reliable in detecting exciting plays. Research in audio-based features have focused on detecting broad events like cheering, music, applause, speech characteristics and employ this information with heuristics to identify exciting plays. Alternatively, emotion analysis of the commentators speech can be a more generic methodology of identifying excitability across a wide-range of games. While some research has used speech-based features (such as mean pitch value in [1]), the possibility remains largely under-explored in sports highlights generation. It is for this reason that we specifically focus on speech-based features for detecting exciting plays. In particular, we employ both spectral and excitation based features such as pitch (F_0), formant frequencies (F_1–F_3), and spectral center of gravity (SCG) which have been shown to work well in stress detection and classification [4, 5, 6]. Our approach also uses a GMM (Gaussian Mixture Model) based classifier to automatically distinguish high and low excitement audio segments. The GMM classifier is trained on human-annotated baseball games where a subjective assessment of the excitement level for different scenes is provided. We use the GMM classifier to assign soft scores to audio segments, which rank orders the segments automatically.

Since the proposed approach is based on speech features, accurate speech background is necessary for good performance. Accurate segmentation in sports videos can be especially challenging due to the low levels of SNR (signal-to-noise ratio) [7] and changes in talking styles [8]. Therefore, existing approaches often rely on supervised audio segmentation algorithms where speech and background models are trained on labeled corpora. However, such an approach is time consuming and often domain dependent. In this study, we circumvent this problem by introducing a new measure of non-stationarity. Interestingly, the new measure is observed to separate a wide range of noise types (and speech) in a reliable and ordered fashion (*i.e.*, increasing order of non-stationarity). Using this new measure, a simple unsupervised algorithm for audio segmentation is proposed. The combination of speech segmentation, excitement measure extraction, and GMM-based excitement level classification constitutes our audio-processing system.

While the audio processing strategy is effective in identifying periods of exciting play, end-points of scenes must be detected to provide meaningful highlights. For this purpose, we use the video signal to detect baseball pitching scenes and slow-motion replay. Detection of these events allows a high-level segmentation of the game play on a pitch-by-pitch basis. Hereafter, pitching scenes are rank-ordered by using the excitement scores of constituent audio segments. This information can now be used to provide highlights of any desirable length.

2. Audio Processing

2.1. Audio-Features Based Segmentation

The proposed segmentation strategy is described below. Let m_{ij} be the Mel-filter bank energy (MFBE) of the ith filter-bank and j^{th} audio-frame. In this study, 40 filter banks are used (*i.e.*, $i =$
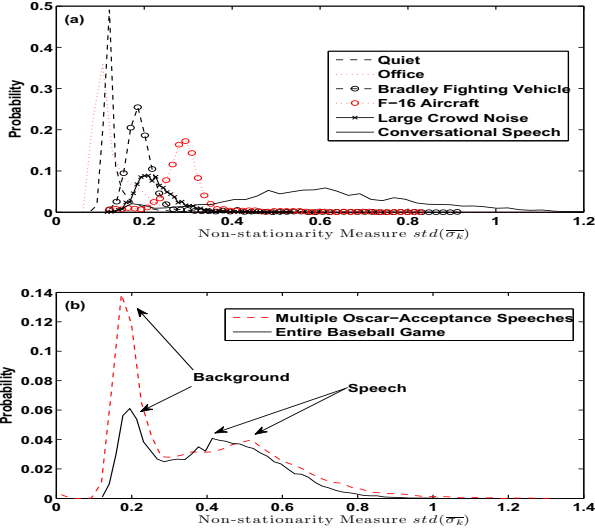
26 – 30 September 2010, Makuhari, Chiba, Japan

Figure 1: *Probability distribution of (a) different unique environments, and (b) mixed environments.*

$1 \ldots 40$) and each audio frame is 25 ms long with 10 ms overlap. Next, the non-stationarity in the signal is estimated by computing the standard deviation of MFBE over a longer time period termed as segments. Let σ_{kj} be the kth standard deviation for the jth Mel-filter bank given by:

$$\sigma_{kj} = \sqrt{\frac{1}{N_s} \sum_{i=(k-1)N_s+1}^{kN_s} (m_{ij} - \frac{1}{N_s} \sum_{i=(k-1)N_s+1}^{kN_s} m_{ij})^2} \quad (1)$$

where N_s is the time period in number of frames. In this study, we choose N_s such that the time period for measuring non-stationarity is 200 ms with a 100 ms overlap.

Our experiments show that the vector $\overline{\sigma_k} = [\sigma_{k1} \ldots \sigma_{k40}]$ as well as the standard-deviation of $\overline{\sigma_k}$ given by $std(\overline{\sigma_k})$ are very effective at distinguishing audio environments. For example, Fig. 1(a) shows the probability distribution function of $std(\overline{\sigma_k})$ for various environment-types, namely, (i) Quiet, (ii) Office, (iii) Bradley Fighting Vehicle, (iv) F-16 Fighter Aircraft, (v) Large Crowd Noise, and (vi) Conversational Speech. The distributions show that the different environments separate effectively within the feature space. For example, Quiet and Office environments show low values of $std(\overline{\sigma_k})$ indicating relatively stationary environments, and Large Crowd and Speech display high values of $std(\overline{\sigma_k})$ indicating highly non-stationary environments. Additionally, Fig. 1(b) shows the distribution of $std(\overline{\sigma_k})$ for (i) Oscar ceremony acceptance speeches, and (ii) commentators speech from baseball games. It is noted that the background for Oscar ceremonies and baseball games contain audio-events like applause, shouting, cheering, whistling, laughing, music, *etc.*. Figure 1 (b) shows the bimodal nature of the non-stationarity measure distribution, with distinct peaks for speech and background. In both scenarios, a suitable threshold can be determined to effectively separate speech and background.

Next, we present a simple unsupervised segmentation algorithm that utilizes the proposed non-stationarity measure for segmentation. First, the non-stationarity measure $std(\overline{\sigma_k})$ is computed for each segment of the entire game video. Next, a 2-mixture GMM is trained using the non-stationarity measure and the expectation-maximization (EM) algorithm. The underlying intuition here is that while one Gaussian would learn speech, the second would learn background distribution characteristics from overall bimodal feature distribution. This learning is now exploited by computing the posterior probability of each mixture

Table 1: Segmentation Accuracy Using the Proposed Technique

	Accuracy	Miss	False-Alarms
Average	80.1%	2.6%	17.3%

component for every feature P_{gk} as:

$$P_{gk} = \frac{1}{\sqrt{2\pi}\sigma_g} \exp(\frac{(std(\overline{\sigma_k}) - \mu_g)^2}{2\sigma_g^2}) \quad (2)$$

where μ_g and σ_g^2 are the mean and variance of the gth Gaussian ($g = 1, 2$). Using the posterior probabilities P_{gk}, each segment can now be assigned to the more likely Gaussian, (*i.e.*, the one with the higher posterior probability). As observed in Fig. 1, the Gaussian with the larger μ_g is more likely to be speech since the non-stationarity of speech is much larger than the typical background acoustics. Using this observation, speech and background Gaussians within the GMM can be identified and every k^{th} segment can be assigned to either speech or background. Speech and background decisions are persistent in time and rapid switching of decisions is very unlikely. This intuition is applied to the algorithm by utilizing Viterbi smoothing to the above decisions while employing a high self-transition probability (values from 0.90 to 0.99 work best). The smoothed decisions are utilized as the final decisions for the remainder of the system. Table 1 shows the segmentation accuracy of the proposed technique using data from 6 separate baseball games (about 15 hours of audio). An accuracy rate of 80.1 % is achieved with very low miss rate of 2.6 % (miss is speech detected as background) and reasonable false-alarm rate of 17.3 % (false-alarm is background detected as speech).

2.2. Speech-based Excitement Analysis

In this section, we search for a set of speech parameters that would be in some way correlated with the excitement level observed in commentators and, hence, would allow for an automatic speech-based spotting of key moments in sports. Past studies have shown that emotions and stress affect a number of speech production parameters [4, 5, 6, 9]. It has been observed that not only speech parameters vary across various emotional and stress classes, but the rate of their change is often proportional to the intensity of the particular emotion or stress.

In the first step, a correlation between selected speech production parameters and human-labeled excitement levels is analyzed. For this purpose, islands of commentators' speech in 6 baseball games were manually labeled by an expert annotator into 4 subjective perceived excitement levels (ordered *level 1* – no excitement, *level 4* – maximum excitement). The following parameters were extracted from the commentators speech in an automatic fashion using WaveSurfer and in-house tools: fundamental frequency F_0, first four formant center frequencies in voiced speech segments F_{1-4}, spectral center of gravity (*SCG*), and so called spectral energy spread (*SES*), which represents a frequency interval of one standard deviation from *SCG*, (*i.e.*, an interval that would capture approximately 34% of the spectral energy, if the spectrum envelope were Gaussian). While in reality the shape of the spectral envelope deviates from Gaussian, we have observed that *SES* provides a reasonable measure of changes in energy spread over frequency and together with *SCG* provides a more noise-robust spectral descriptor than spectral slope [6].

Figures 2 and 3 show the distribution of mean F_0 and *SCG* across human labeled excitement levels and games (error bars denote 95% confidence intervals). It can be seen that while the range of parameter values varies across games, due to the varying physiological properties and talking manners of the actual commentators, there is an increasing trend in F_0 and *SCG* with the level of perceived excitement. Similar observations were made for F_{1-3} and *SES*. To assess the degree of correlation between the speech
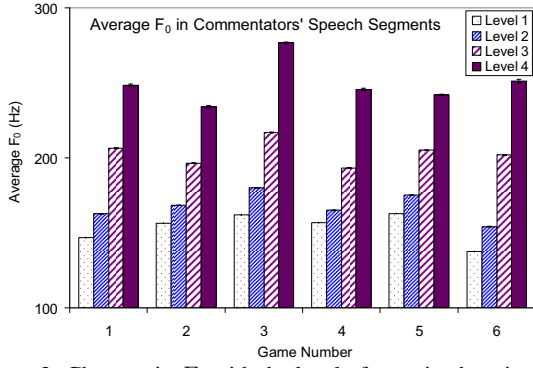
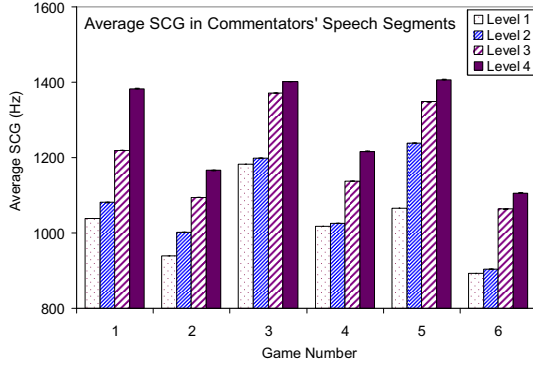Figure 2: Changes in F_0 with the level of perceived excitement.

Figure 3: Changes in SCG with level of perceived excitement.

Figure 4: Linear regression - mean/variance normalized F_0.

Figure 5: Linear regression - mean/variance normalized SCG.

parameters and perceived excitement levels, a linear regression was conducted for all parameters. To compensate for the inter-commentator differences across games, all parameters were normalized to zero mean and unity variance at the game level, by subtracting a game-dependent parameter mean from all respective game samples, and dividing them by a game-dependent standard deviation. We note that this type of normalization assumes an off-line processing of the game recording. The outcomes of linear regression are shown for F_0 and SCG in Fig. 4 and 5, and summarized for all parameters in Table 2.

The degree of linear relationship between the subjective excitement levels and parameter changes are represented by the correlation coefficient R^2. The spread of the actual samples around the estimated regression line is measured by the means of mean square error (MSE). It can be seen in Table 2 that mean game F_0, SCG, and F_{1-2} exhibit a relatively high linear relationship with subjective excitement labels, while F_3 and SES display just a moderate relationship (also note increased MSE values), and F_4 seems to be unaffected by the perceived excitement.

Based on the correlation analysis, F_0, SCG, and F_{1-3} were selected to form a feature vector for automatic excitement-level assessment. A Gaussian Mixture Model (GMM) maximum likelihood (ML) classifier was trained on the feature vectors extracted from 4 baseball games, utilizing the subjective excitement levels as transcriptions of the training data, and evaluated on 2 distinct games representing the open test set. The task was to distinguish 'moderate' excitement (corresponding to subjective excitement levels 1–2) and 'high' excitement (levels 3–4). During the test phase, a binary decision threshold yielding an equal error rate (EER) was searched in an iterative procedure. To evaluate the repeatability of the results, the experiment was repeated 3x in a round robin scheme. In all cases, 4 index-wise adjacent games were used for training and two games for testing. The overall excitement level classification results for islands of commentators speech are shown in Table 3, accompanied by the confusion matrices ('Mod' stands for moderate excitement). It can be seen that
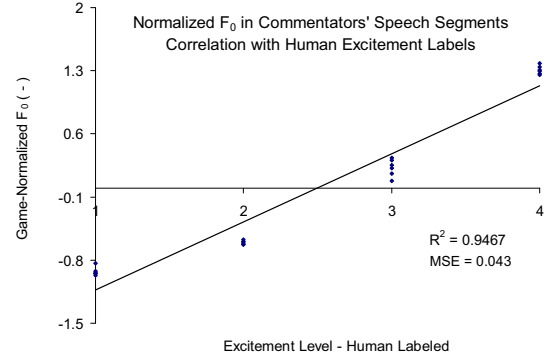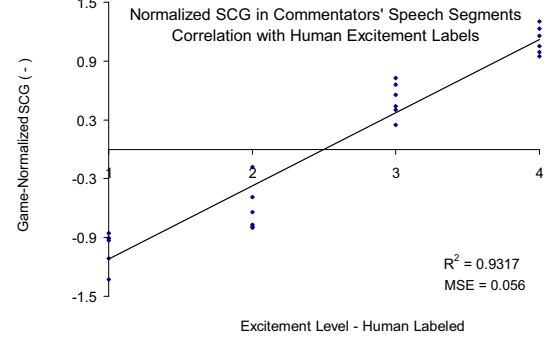
the EER in the round robin scheme range from 21.4–22.4 %. It is noted that the binary decision threshold in the ML classifier can be adjusted to reduce the probability of missed high excitement islands, at the costs of increased probability of false alarms from the moderate excitement islands.

3. Video processing

3.1. Video Shot Boundary Detection

First, the video is segmented using the cut detection method presented in [10]. A 48 dimensional color histogram based features extracted from each video frame are used for this purpose. For each color, we subdivide the color range into 16 equal intervals and compute the number of pixels in that range. Thus, 16 coefficients of the feature is generated for each color yielding a dimension of 48.

3.2. Pitching shot detection

Baseball pitching shots are detected based on the approach presented in [11]. Grass and soil color pixels were detected using their respective color distribution in the HSV color space [11]. The area ratio [12], R_a is then computed and used to classify the shot in three categories based on the rules: (i) If $R_a > 45\%$ then it is an outfield scene, (ii) If $25\% < R_a < 45\%$ then a pitching scene, and (iii) If $R_a < 25\%$ then other scene.

We ensure that our miss rate is a minimal in the first stage. For each frame i classified as a pitching scene from the first pass, three binary conditions, $C_1(i)$, $C_2(i)$ and $C_3(i)$ are set in the following manner. The default value of these variables is set to FALSE.

- $C_1(i)$: If the number of field pixels in the lower half of the frame is more than 2 times greater than that of the higher half, $C_1(i)$ =TRUE.
- $C_2(i)$: Compute the vertical profile of the field pixels and search for a valley. If there is a strong valley on the left side of the screen such that the value is less than the mean value of the average profile, $C_2(i)$ =TRUE.

Table 2: Correlation analysis.

	F_0	F_1	F_2	F_3	F_4	SCG	SES
R^2	0.947	0.926	0.922	0.779	0.018	0.932	0.538
MSE	0.043	0.081	0.063	0.181	0.803	0.056	0.378

Table 3: Excitement level classification; equal error rates (%).

	Round Robin					
	1		2		3	
Ground Truth	Mod	High	Mod	High	Mod	High
Mod	1579	431	1972	536	2558	738
High	83	304	123	444	171	597
EER (%)	21.4		21.6		22.4	

- $C_3(i)$: Compute a binary edge image of the current frame. The frame is divided into 16 equal blocks [12] and the edge image is analyzed in each block to determine the presence of the pitcher and the batter. If the image intensity in blocks 7, 10, 11, and 14 is greater than the average intensity of the image, $C_3(i)$ =TRUE.

We declare the frame as a pitching shot if for a frame i, the boolean variable $P = C_1(i) \cdot (C_2(i) + C_3(i))$ yields a TRUE value, where $+$ and $\cdot$ indicate the binary OR and AND operation.

3.3. Slow motion detection

We utilized the pixel-wise mean square distance (PWMSD) features for detecting the slow motion regions. Slow motion fields are usually generated by frame repetition or drop, which cause frequent and strong fluctuations in the PWMSD features, $D(t)$. This fluctuation can be measured using a zero crossing detector as described in [13]. First, the $D(t)$ feature is segmented in small windows of N frames. In each window, the zero crossing detection is performed and if it is greater than some predefined threshold, λ the window is assumed to contain slow motion frames.

4. Automatic Highlights Generation

The proposed system is summarized with the major components shown in Fig. 6. For automatic highlights generation, each pitching scene end-points are first determined using the technique described previously. The scene end-points provide a high-level play by play segmentation of the game. Next, the excitement level for each of these scenes is determined by using the GMM-based excitement classifier described previously. It is noted that the log-likelihood ratio of the GMM classifier itself is used as a soft score to represent the level of excitement. Based on these score assignments, the pitching scenes are rank-ordered by excitement level. Now, automatic highlights can be generated by combining the top N exciting scenes, where N is determined based on user-specified time length. Some ex-
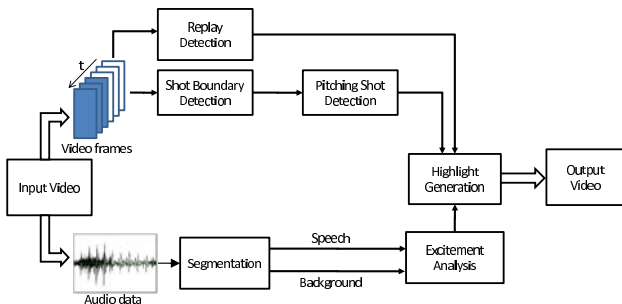
amples of highlights can be viewed on the following website: *http://crss.utdallas.edu/demos/highlights.html*.

5. Conclusion

In this study, a novel methodology that uses estimates of excitability in sports video to create automatic highlights was presented. The new method uses speech-based emotion/stress features to estimate excitement in baseball videos. In this manner, it complements existing approaches that rely on video or audio based features to detect excitement. Furthermore, a novel unsupervised audio segmentation technique that separates speech from background in noisy sports videos was also presented. The new technique uses a measure of non-stationarity to identify and separate disparate environment types. Additionally, video-processing techniques were employed to detect pitching and slow-motion scenes in order to identify end-points of plays more effectively. Finally, the combination of segmentation, excitement-estimation, and scene-identification was uses to create automatic game highlights. The techniques presented in this study are generic and may be equally applicable to a variety of domains.

6. REFERENCES

[1] C. Liu, Q. Huang, S. Jiang, L. Xing, Q. Ye, and W. Gao, "A framework for flexible summarization of racquet sports video using multiple modalities," *Computer Vision and Image Understanding*, vol. 113, pp. 415–424, 2009.

[2] E. Kijak, G. Gravier, P. Gros, L. Oisel, and F. Bimbot, "HMM based structuring of tennis videos using visual and audio cues," in *ICME*, 2003.

[3] R. Radhakrishnan, Z. Xiong, A. Divakaran, and Y. Ishikawa, "Generation of sports highlights using a combination of supervised and unsupervised learning in audio domain," in *Pacific Rim Conference on Multimedia*, 2003.

[4] John H. L. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech Comm.*, vol. 20, no. 1-2, pp. 151–173, 1996.

[5] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, Jan. 2001.

[6] H. Bořil, T. Kleinschmidt, P. Boyraz, and J. H. L. Hansen, "Impact of cognitive load and frustration on drivers' speech," *The Journal of the Acoust. Soc. of America*, vol. 127, no. 3, pp. 1996–1996, 2010.

[7] H. K. Maganti, P. Motlicek, and D. Gatica-Perez, "Unsupervised speech/non-speech detection for automatic speech recognition in meeting rooms," in *Proc. ICASSP'07*, Honolulu, Hawaii, USA, 2007, pp. 1037–1040.

[8] J. Volín, R. Skarnitzl, and P. Pollák, "Confronting HMM-based phone labelling with human evaluation of speech production," in *Proc. INTERSPEECH'05*, Lisbon, Portugal, 2005, pp. 1541–1544.

[9] Z. Callejas and R. López-Cózar, "Influence of contextual information in emotion annotation for spoken dialogue systems," *Speech Communication*, vol. 50, no. 5, pp. 416 – 433, 2008.

[10] B.T. Truong, C. Dorai, and S. Venkatesh, "New enhancements to cut, fade, and dissolve detection processes in video segmentation," in *Proceedings of the eighth ACM international conference on Multimedia*. ACM, 2000, p. 227.

[11] W.T. Chu and J.L. Wu, "Explicit semantic events detection and development of realistic applications for broadcasting baseball videos," *Multimedia Tools and Applications*, vol. 38, no. 1, pp. 27–50, 2008.

[12] C.C. Lien, C.L. Chiang, and C.H. Lee, "Scene-based event detection for baseball videos," *Journal of Visual Communication and Image Representation*, vol. 18, no. 1, pp. 1–14, 2007.

[13] H. Pan, P. van Beek, and M. I. Sezan, "Detection of slow-motion replay segments in sports video for highlights generation," in *ICASSP-01*, pp. 1649–1652.

Figure 6: The highlight generation system block diagram