

RESEARCH

Open Access

# Multi-modal highlight generation for sports videos using an information-theoretic excitability measure

Taufiq Hasan, Hynek Bořil, Abhijeet Sangwan and John H L Hansen\*

## Abstract

The ability to detect and organize ‘hot spots’ representing areas of excitement within video streams is a challenging research problem when techniques rely exclusively on video content. A generic method for sports video highlight selection is presented in this study which leverages both video/image structure as well as audio/speech properties. Processing begins where the video is partitioned into small segments and several multi-modal features are extracted from each segment. Excitability is computed based on the likelihood of the segmental features residing in certain regions of their joint probability density function space which are considered both exciting and rare. The proposed measure is used to rank order the partitioned segments to compress the overall video sequence and produce a contiguous set of highlights. Experiments are performed on baseball videos based on signal processing advancements for excitement assessment in the commentators’ speech, audio energy, slow motion replay, scene cut density, and motion activity as features. Detailed analysis on correlation between user excitability and various speech production parameters is conducted and an effective scheme is designed to estimate the excitement level of commentator’s speech from the sports videos. Subjective evaluation of excitability and ranking of video segments demonstrate a higher correlation with the proposed measure compared to well-established techniques indicating the effectiveness of the overall approach.

## 1 Introduction

Automatic video analysis and summarization has a wide range of applications in domains such as sports, movies, security, news and on-line video streaming. Hot-spot information can be utilized in technologies such as search, summarization, and mash-ups, in addition to navigation of multimedia content. For example, emotional ‘hot-spots’ within sports videos are usually more exciting than the overall game video, which motivates the formulation of a solution to automatically generate highlights from such videos. Various approaches towards automatic event detection and summarization in sports videos have been presented in the literature. Past methods utilize information from a single modality [1], or

combine multiple modalities in different ways [2-7]. Many techniques depend on specific sports type [2,8,9], video effects [1], or environments. Methods such as those used by Lein et al. [8] depend on annotating the full game automatically using sophisticated machine learning and domain knowledge, whereas other methods tend to be more generic [3,10-12]. In simpler methods such as in [13] as applied to baseball games, the probability of a baseball hit and excited speech is combined to estimate the excitability of a video segment. In [10], a generic approach was presented to estimate expected variations in a user’s excitement from the temporal characteristics of selected audio-visual features and the editing scheme of a video. In general, generic highlight extraction schemes aim at constructing temporal features from audio/video streams that are proportional to or indicate user excitability [14]. Later, some kind of fusion strategy is used to generate a single excitement curve [10] providing estimated affective state of the viewer at different points in time/video segment.

\*Correspondence: john.hansen@utdallas.edu  
Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering and Computer Science, Department of Electrical Engineering, University of Texas at Dallas, 2601 N. Floyd Road, EC33, Richardson, TX 75080-1407, USA

In our initial study [15], we examined a simple audio/video feature fusion method for baseball highlight extraction. In this paper, we extend the feature space and propose an information-theoretic measure of excitability for sports highlight selection in a generic framework [16]. Our proposed measure is based on a simple but powerful principle of information theory: the less likely an event, the more information it contains. We assume that interesting parts in a video occur rarely [4] and therefore have high self-information (also known as the ‘surprisal’) [17]. This can be intuitively understood as follows: if for a given sports video the ambient crowd noise is always high, then audio energy as an excitement indicator [10] would be unreliable, i.e., in this game, there is nothing ‘surprising’ in high audio energy. Our proposed method aims at estimating the user excitability directly from low-level features using their joint-PDF estimated over the game videos. Even when extended videos are not available for training these models, the proposed technique can still extract highlights from a given game video by estimating the feature PDFs from itself in an unsupervised fashion, provided that the features used are generally related to user excitement. An advantage of the proposed method is that it is less affected by extreme values of a single feature due to off-field distractions [5] since the joint behavior of the features is considered in a probabilistic framework. Using the proposed excitability measure, the video segments can be rank-ordered to automatically generate highlights. The technique can also be used to estimate an excitement-time curve [10] to demonstrate user-affective states over a time sequence of the video stream.

The multi-modal events/features used in the proposed highlights generation system are the following: slow motion replay, camera motion activity, scene cut density, and excitement in commentators’ speech. Past studies [5,6] typically consider simplistic features such as energy, zero crossing rate, and others to estimate excitement from the audio modality. Inspired by studies on emotion assessment [18-20], in this study, we analyze the effects of excitement on the parameters of a linear model of speech production derived from commentators’ speech. As will be shown, some speech parameters are strongly correlated with the perceptual excitability and hence, are selected to form a feature vector for the audio-based excitement assessment.

This paper is organized as follows: section 2 proposes the probabilistic excitability measure and discusses methods of its implementation in highlight selection. In section 3, analysis of speech production features and their correlation with excitement in commentators’ speech in the context of sports videos is discussed. In section 4, the overall highlights extraction scheme is presented; section 5 details a subjective evaluation of the full system and discusses results, and section 6 concludes the study.

## 2 Proposed excitability measure

At first, the video is divided into small segments for feature extraction. Next, several features (scalar parameters) are extracted from each segment that are modeled to be generally proportional to the user’s excitement of the given segment. These features represent long/short term (cumulative) characteristics from different modalities, such as duration of excited speech, average motion activity, and others.

### 2.1 Basic formulation

Let the random variable  $X_i$  be the  $i$ th feature and  $x_i(k)$  be an observation of that feature in the  $k$ th segment. Since  $X_i$  is in general proportional to the excitability of the video segment,  $p(X_i \geq x_i(k))$  will be very low for highly exciting video segments, (i.e., they will be rare outcomes for the random event  $\{X_i \geq x_i(k)\}$ ). Therefore, the self-information measure (in bits) associated with the random event  $\{X_i \geq x_i(k)\}$  given by [17]

$$\begin{aligned} \zeta_i(k) &= I[X_i \geq x_i(k)] \\ &= -\log_2 [p(X_i \geq x_i(k))] \end{aligned} \quad (1)$$

will be proportional to excitability. For  $D$  feature parameters, we define the random vector  $\mathbf{X} = (X_1, X_2, \dots, X_D)$  as the feature vector, and  $\mathbf{x}(k)$  as an observation vector in the  $k$ th video segment. We can now refine Equation 1 for  $D$  dimensions as

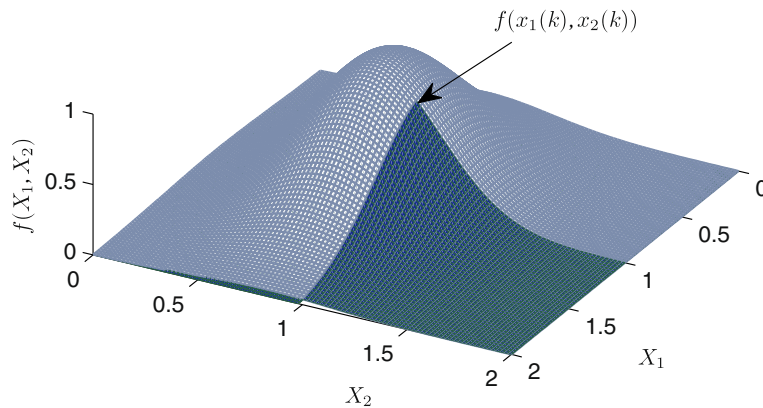
$$\zeta(k) = -\log_2 [p(\mathbf{X} \geq \mathbf{x}(k))], \quad (2)$$

where  $\zeta(k)$  is a measure of excitability for segment  $k$  from  $D$  features. Assuming that  $X_1, X_2, \dots, X_D$  are independent, we have

$$\begin{aligned} \zeta(k) &= -\log_2 \prod_{i=1}^D p(X_i \geq x_i(k)) \\ &= -\log_2 \prod_{i=1}^D \int_{x_i(k)}^{\infty} f_{X_i}(\lambda) d\lambda, \end{aligned}$$

where  $f_{X_i}$  is the PDF of the  $i$ th feature. The idea is illustrated in Figure 1 with two features  $X_1$  and  $X_2$ . For an observation,  $\mathbf{x}(k) = (x_1(k), x_2(k))$  obtained from the  $k$ th segment, the area under the shaded region, determines how likely it is that other segments would have higher feature values compared to this observation.

The advantage of using the proposed measure is that it not only considers the value of the observation  $x_i(k)$  in the  $k$ th segment, but also takes into account how likely it is that this feature yields a higher value than  $x_i(k)$ . Thus,  $\zeta(k)$  can be used to rank video segments from a high to low excitement level.



**Figure 1** Conceptual joint PDF of two multimodal features  $X_1$  and  $X_2$  extracted from video segments. Shaded area reveals the high tail region indicating exciting events.

### 2.2 Incorporating feature reliability

In many applications, some feature parameter is more reliable/accurate than others. In the proposed scheme, reliability or relative importance of different features can be easily incorporated. We introduce the weight parameters  $\eta_i$  for the  $i$ th feature as follows:

$$\zeta(k) = -\log_2 \prod_{i=1}^D [p(X_i \geq x_i(k))^{\eta_i}], \quad (3)$$

where  $\sum_{i=1}^D \eta_i = 1$  and  $\forall i : 0 \leq \eta_i < 1$ . If it is known *a priori* that some features are more reliable than others, then appropriate weights can be set. On the other hand, the correlation of the individual feature parameters to the subjective excitability can be obtained on a development dataset. These correlation values will give indications on which features are more reliable, i.e. more related to user excitement, and thus be weighted higher. We will discuss this further in the experiments section.

### 3 Excitement measurement in speech

Most current highlight extraction methods utilizing game audio tend to focus on simplistic features such as audio energy or short-time zero crossing [5,6] to estimate the excitement level. Past literature on emotions and stress suggests that a number of speech production parameters can be affected by varying speech modalities [18-21]. In this section, we extract a set of speech parameters derived from the linear model of speech production and evaluate their correlation with the perceptual excitement level in the commentators audio. The subset of parameters displaying a strong correlation with the excitement level is identified and used as features in the design of an audio-based excitement classifier<sup>a</sup>.

For the purpose of the correlation analysis and subsequent classifier evaluations, islands of commentators speech in six baseball games were manually labeled by an

expert annotator into four perceptual excitement levels (ordered from *level 1* – no excitement, to *level 4* – maximum excitement). WaveSurfer [22] and in-house tools were used to extract the following speech production parameters from the speech segments: fundamental frequency  $F_0$ , first four formant center frequencies in voiced speech segments  $F_{1-4}$ , spectral center of gravity (SCG),

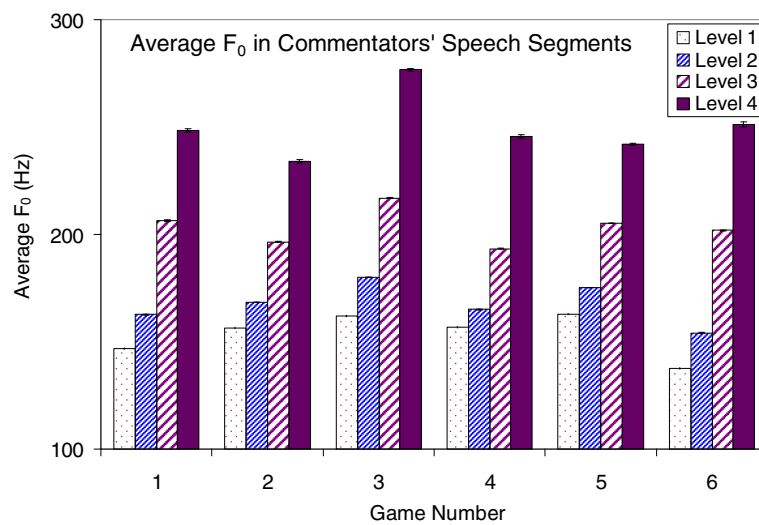
$$SCG = \frac{\sum_{k=1}^N X(k) \cdot k}{\sum_{k=1}^N X(k)}, \quad (4)$$

where  $X(k)$  is the  $k$ th bin in the energy spectrum and  $N$  is the window length; and a so called spectral energy spread (SES),

$$SES = \sqrt{\frac{\sum_{k=1}^N X(k) \times (k - SCG)^2}{\sum_{k=1}^N X(k)}}, \quad (5)$$

which represents a frequency interval of one standard deviation in the distribution of energy spectrum with the mean equal to SCG. We have observed that SES, when combined with SCG, constitutes a more noise-robust spectral descriptor for emotion and stress classification than spectral slope [20].

The distribution of game-specific means of  $F_0$ ,  $F_1$ ,  $F_2$ , and SCG across the four perceptual excitement levels is summarized in Figures 2, 3, 4, and 5. It can be seen that the range of parameter values varies for different games due to the unique physiological characteristics and talking manners of the different game commentators. However, these parameters display in general an increasing trend with the level of excitement. Similar observations were made for  $F_3$  and SES. To assess the

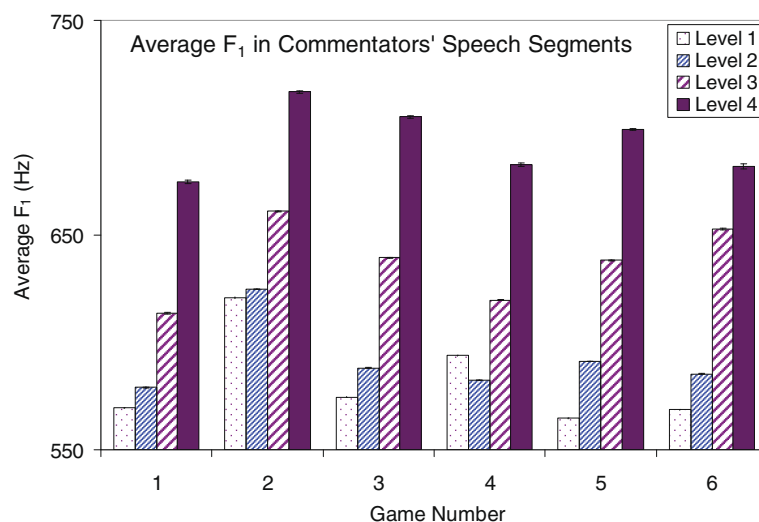


**Figure 2** Changes in  $F_0$  with the level of perceived excitement.

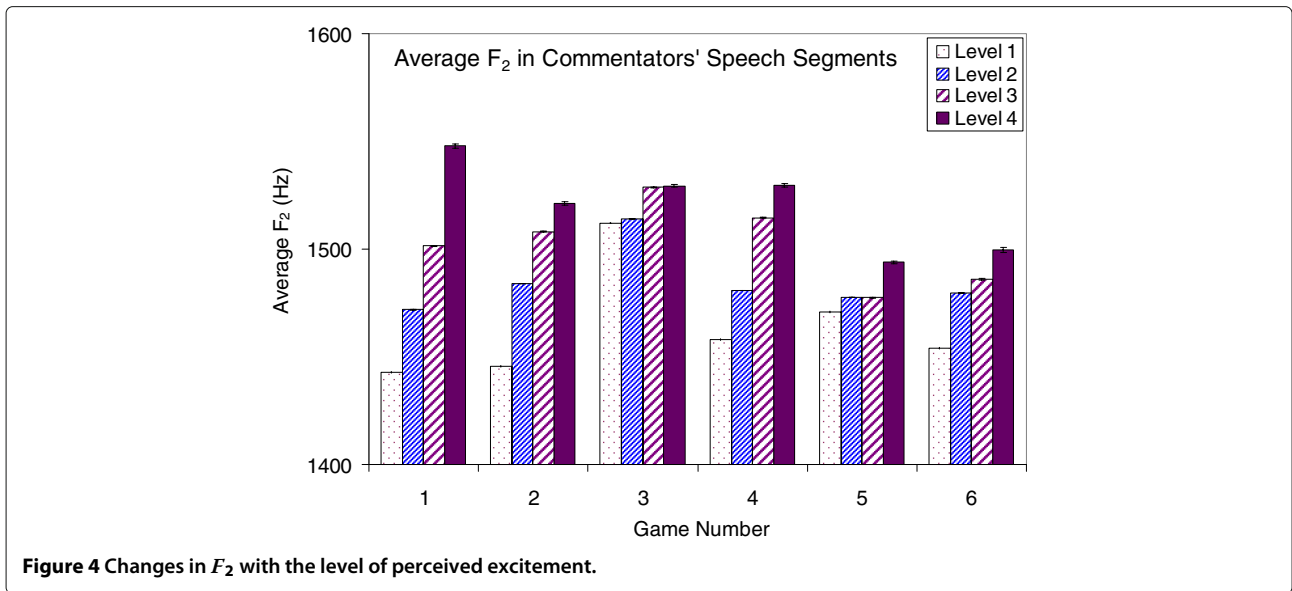
degree of correlation between the speech parameters and perceived excitement levels, a linear regression was conducted. To compensate for the inter-commentator differences across games, all parameter distributions were normalized to a zero mean and unit variance at the game level by subtracting a game-dependent parameter mean from all respective game samples, and dividing them by a game-dependent standard deviation. We note that this type of normalization assumes availability of all game audio samples prior to the game highlights generation. Arguably, it would be difficult or even impossible to select most exciting segments prior to 'seeing' the game as a whole, hence this assumption seems quite reasonable. However, if an on-line feature extraction were

preferable in some applications, a cumulative estimation of the commentators mean and standard deviation statistics could be performed on-the-fly. We provide more discussion on this in section 5.1. The outcomes of linear regression are shown for  $F_0$ ,  $F_1$ ,  $F_2$ , and SCG in Figures 6, 7, 8, and 9 and summarized for all analyzed parameters in Table 1.

Table 1 suggests 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 have just a moderate relationship (also note increased MSE values), and  $F_4$  is almost unaffected by the perceived excitement. This corresponds well with the observations made in the past literature. Variations of vocal effort, typical for excited speech,



**Figure 3** Changes in  $F_1$  with the level of perceived excitement.



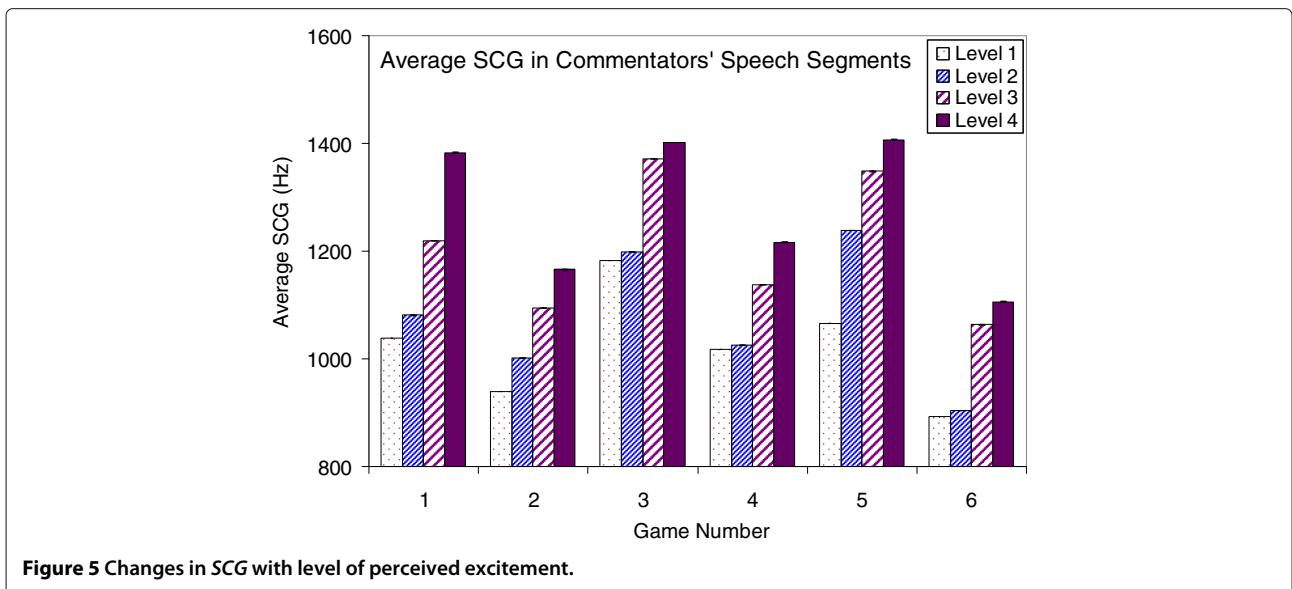
are carried out by both varying sub-glottal pressure and tension in the laryngeal musculature [23]. Pitch (in log frequency) changes almost linearly with vocal intensity [24]. In the spectral domain, the energy in increased vocal effort speech migrates to higher frequencies, causing an upward shift of SCG [25], and flattening of the spectral slope of short-time speech spectra [18,26].  $F_1$  is inversely proportional to the vertical position of the tongue and  $F_2$  rises with tongue advancement [27]. The increased vocal effort in excited speech is likely to be accompanied by a wider mouth opening, which is realized by lowering the jaw and the tongue. As a result,  $F_1$  will increase in frequency [23,28].  $F_2$  rises in some phones [29] while may decrease in others [30]. On the other hand, locations of

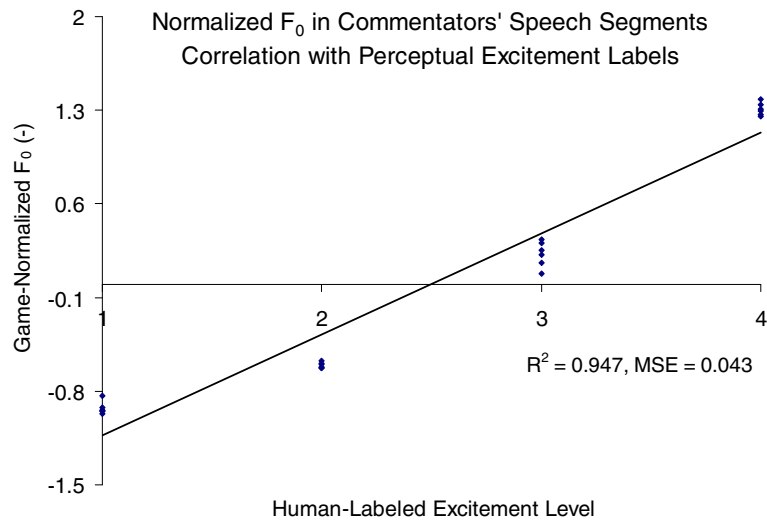
higher formants are rather determined by the vocal tract length [31] and as such are not as sensitive to the vocal effort variations.

Based on the results in Table 1,  $F_0$ , SCG, and  $F_{1-3}$  are chosen as features for the automatic excitement-level assessment. The excitement level classification is conducted using a Gaussian mixture model (GMM)-based classifier and will be discussed in more detail in section 5.1.

#### 4 Highlights extraction system

We use six baseball game videos from the 1975 World Series to evaluate the proposed highlight generation method. The collected videos are of resolution  $720 \times 480$



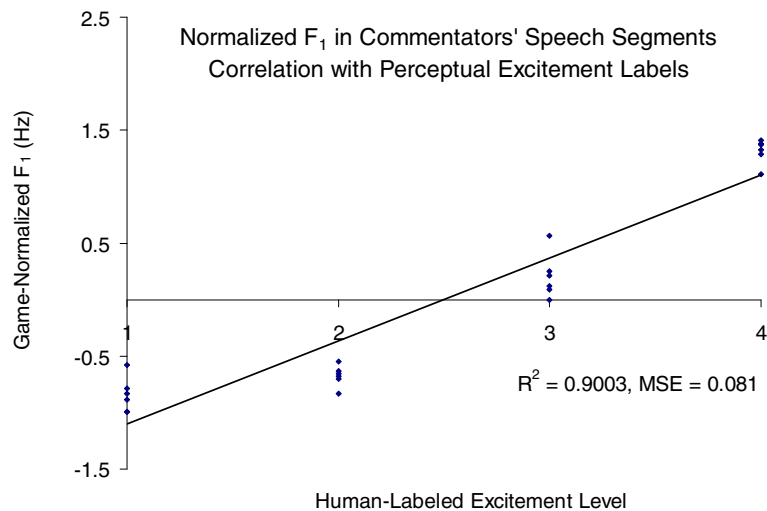


**Figure 6** Linear regression - mean/variance normalized  $F_0$ .

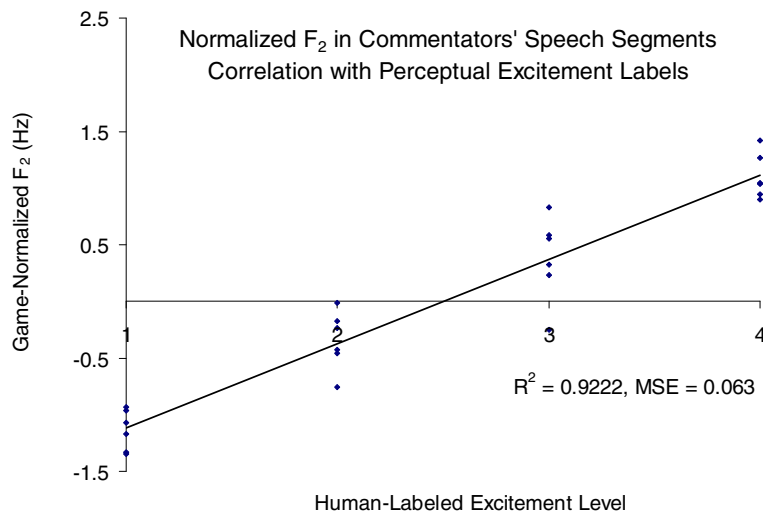
pixels. A block diagram of the overall system is given in Figure 10. Our highlights video generation depends on a *semantic video segmentation*, though other method of segmentation can also be utilized. We define *semantic segments* as short self-explanatory video segments that can be used as building blocks for the highlights video. Examples of such segments can be play times in soccer games, time interval between each bowling in cricket, times between each pitching in baseball games, etc. For our experiments, we perform segmentation at the pitching scenes. This is the only part of the highlights generation process which is game dependent. Later in section 4.4, we demonstrate how the proposed measure can also be used

to analyze continuous excitement-time characteristics of a sports video.

The notation used from this point forward is as follows:  $t$ ,  $k$ , and  $i$  denote video frame, video segment, and feature index, respectively. For the  $i$ th feature,  $\Phi_i(t)$ ,  $x_i(k)$ , and  $G_i(t)$  indicate feature value at time  $t$ , feature parameter extracted from segment  $k$ , and viewer arousal curve at time  $t$  estimated as in [10], respectively. The multi-modal events/features used for excitability measure include: (1) slow motion replay, (2) camera motion activity, (3) scene cut density, (4) commentators' speech in high and (5) low excitement levels, and (6) audio energy. For comparison, we also implemented the highlight selection method



**Figure 7** Linear regression - mean/variance normalized  $F_1$ .



**Figure 8** Linear regression - mean/variance normalized  $F_2$ .

presented in [10]. The details of the system are discussed below.

#### 4.1 Video processing

##### 4.1.1 Slow motion detection

A pixel-wise mean square distance (PWMSD) feature is used for detecting slow motion regions [1]. Slow motion fields are usually generated by frame repetition or drops, which cause frequent and strong fluctuations in the PWMSD features. This fluctuation can be measured using a zero crossing detector as described in [1]. First, the PWMSD feature stream  $D(t)$  is segmented into small windows of  $N$  video frames. In each window, the zero crossing detection is performed,

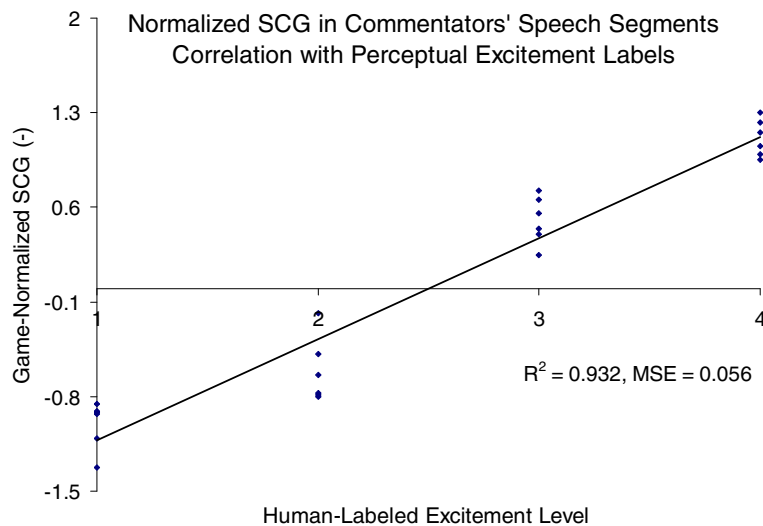
$$Z_c(t, \theta) = \sum_{j=1}^{N-1} \Psi(D(t-j) - \bar{D}(t), D(t-j-1) - \bar{D}(t), \theta) \quad (6)$$

where  $\bar{D}(t)$  is the mean value of  $D(t)$  in the sliding window at time  $t$ , and

$$\Psi(x, y, \theta) = \begin{cases} 1 & \text{if } x \geq \theta \text{ \& } y \leq \theta \text{ or } x \leq -\theta \text{ \& } y \geq \theta \\ 0 & \text{otherwise.} \end{cases} \quad (7)$$

Next, the  $Z_c()$  function outputs for each window is considered and if it is greater than some predefined threshold  $\lambda$ , the window is assumed to contain slow motion frames. We use  $\lambda = 15$ .

Since slow motion replay is displayed after some interesting events in sports, we assume that the duration of



**Figure 9** Linear regression - mean/variance normalized SCG.

**Table 1 Correlation analysis**

	$F_0$	$F_1$	$F_2$	$F_3$	$F_4$	SCG	SES
R2	0.947	0.900	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

a slow motion shot in the  $k$ th semantic segment is proportional to excitability (given the segment is sufficiently long) and thus we use this measure as the feature parameter  $x_1(k)$ . To obtain  $G_1(t)$ , we first define the slow motion function as,  $\Phi_1(t) = 1$  if slow motion is detected at time  $t$ ; or 0 otherwise (Figure 11a). Next, we filter  $\Phi_1(t)$  to obtain  $G_1(t)$  to fulfill the ‘smoothness’ criteria required for the method presented in [10]. In general, for the  $i$ th feature, we use the following filter:

$$G_i(t) = \Phi_i(t) * K_t(l, \beta), \quad (8)$$

where  $K_t(l, \beta)$  indicates a Kaiser window [32] of length  $l$  and scale parameter  $\beta$  ( $l = 500$  and  $\beta = 5$  is used) given by,

$$K_t(l, \beta) = \begin{cases} \frac{I_0\left(\pi\beta\sqrt{1-\left(\frac{2t}{l}-1\right)^2}\right)}{I_0(\pi\beta)}, & 0 \leq t \leq l-1 \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

#### 4.1.2 Camera motion estimation

In sports videos, high motion in the camera usually indicate exciting events [33]. For detecting camera motion, we use a block-matching algorithm [34] to estimate the motion vector between successive video frames. A large (64 by 64 pixels) block size is used in order to reduce

the motion estimation sensitivity to movement of small objects within the frame. The raw motion values are normalized and stored in  $\Phi_2(t)$ , then smoothed using Equation 8 to obtain  $G_2(t)$ .

We observe that the amplitude of the resultant motion vector calculated in each frame gives a good indication of camera movement, such as pan and zoom. Thus, segmental feature  $x_2(k)$  is computed by averaging  $G_2(t)$  across the  $k$ th segment.

#### 4.1.3 Scene-cut density

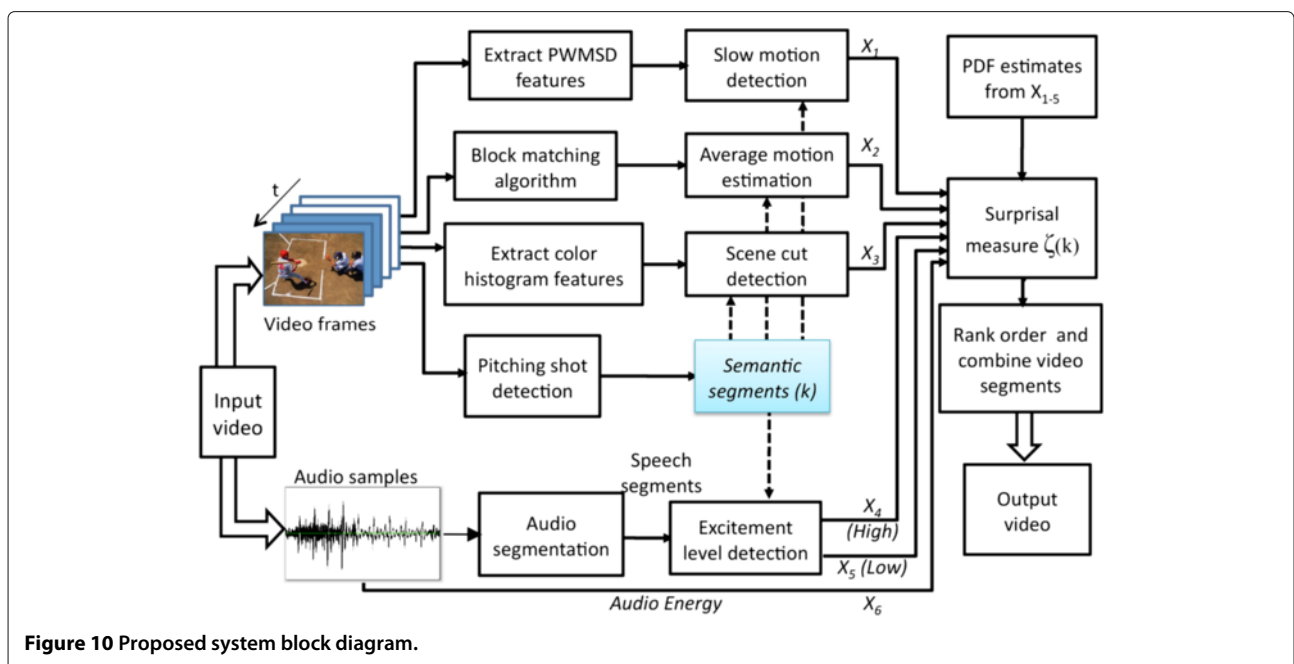
We utilize the cut detection method proposed in [35]. A 48-dimensional color histogram-based feature is used for this purpose. In [10], it is shown that scene cut density measure, which analyzes the influence of shot duration in user excitability, is correlated with excitement in sports videos. This measure is used in our scheme and is extracted as follows. At each video frame  $t$ , we compute

$$\Phi_3(t) = e^{(1-n(t)-p(t))/\delta}, \quad (10)$$

where  $n(t)$  and  $p(t)$  are frame indices of the two nearest scene-cuts to the left and right of the frame  $t$ , respectively. The parameter  $\delta$  is set to 500. Again, we use (8) to obtain  $G_3(t)$  from  $\Phi_3(t)$  and average  $G_3(t)$  over the  $k$ th segment to compute  $x_3(k)$ .

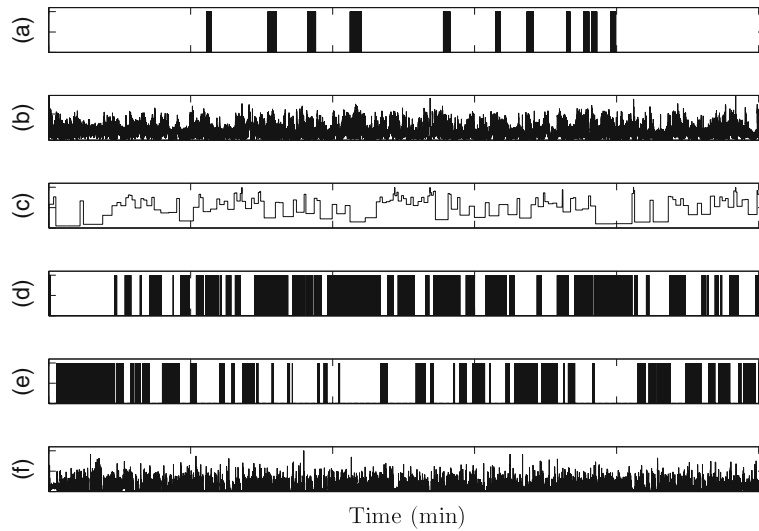
#### 4.1.4 Pitching scene detection

The times when the pitching takes place are very well suited locations for semantic segmentation in baseball



**Figure 10** Proposed system block diagram.

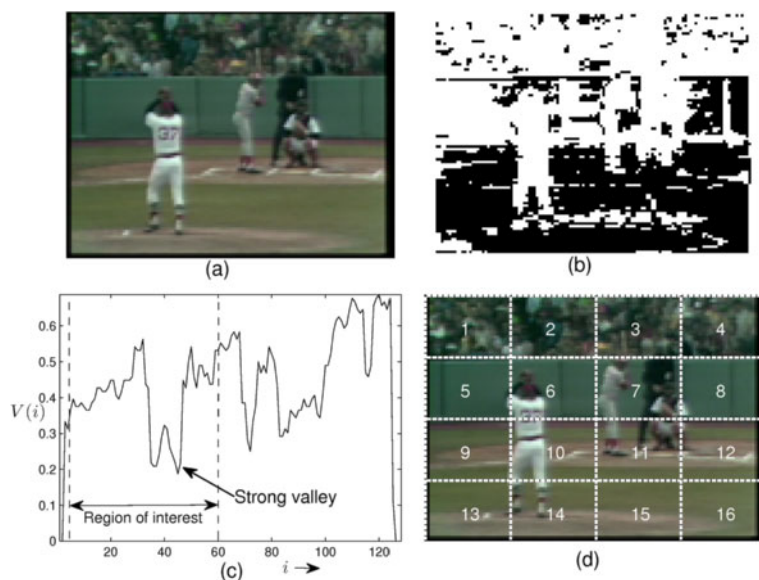




**Figure 11** A timeline view of the detected events/feature functions  $\Phi_i(t)$  segment of a baseball game video. **(a)** Slow motion function  $\Phi_1(t)$ , **(b)** motion activity function  $\Phi_2(t)$ , **(c)** scene-cut density  $\Phi_3(t)$ , **(d)** high excitement regions in speech  $\Phi_4(t)$ , **(e)** low excitement regions in speech  $\Phi_5(t)$ , and **(f)** Audio energy  $\Phi_6(t)$ .

games. To detect the pitching scenes, the following operations are performed on each video frame: the field pixels are detected using the HSV color space condition as in [36] and a  $H \times W$  binary image is formed. Here,  $W$  and  $H$  denote image width and height in pixels, respectively. Figure 12b shows an example binary image from a pitching scene. We test four conditions that can be an indication of a pitching scene, and four boolean variables  $C_A$ ,  $C_L$ ,  $C_V$ , and  $C_P$  are set as follows:

- i) *Area ratio condition ( $C_A$ )*: Area ratio [36],  $R_a$ , is computed from the binary image  $I_f(\cdot, \cdot)$ . If  $25\% \leq R_a \leq 45\%$  then  $C_A = 1$ ; else 0.
- ii) *Layout condition ( $C_L$ )*: In pitching scenes, the lower half of the image usually contains more baseball field pixels [36]. Thus, if the lower half has more than twice the number of field pixels compared to the upper half, we set  $C_L = 1$ , else 0, which becomes:



**Figure 12** Pitching scene detection. **(a)** a sample pitching scene, **(b)** detected field pixels (shown in black), **(c)** vertical profile  $V(i)$  of the field pixels, and **(d)** 16 blocks dividing the image.

if

$$\sum_{i=1}^W \sum_{j=1}^{H/2} I_f(i, j) \geq 2 \sum_{i=1}^W \sum_{j=H/2+1}^H I_f(i, j)$$

then  $C_L = 1$ , otherwise  $C_L = 0$ .

- iii) *Vertical profile condition ( $C_V$ )*: The vertical distribution of the field pixels  $V(i)$  is given by the equation

$$V(i) = \frac{1}{H} \sum_{j=1}^H I_f(i, j). \quad (11)$$

In pitching scenes, a strong valley is usually found in the left hand side of this distribution [36], as shown in Figure 12c, due to the presence of the pitcher. If such a valley is found,  $C_V = 1$ ; else 0.

- iv) *Player location condition ( $C_P$ )*: From the video frame, a binary edge image is calculated using the ‘sobel’ method and image dilation [37] is performed. The resulting image is then divided into 16 equal blocks as shown in Figure 12d. In pitching scenes, a higher intensity in the edge-dilated image will be observed in blocks 7, 10, 11, and 14 due to the presence of the pitcher and the batter [8]. If the image intensity of these regions is greater than the average intensity of the image,  $C_P = 1$ ; else 0.

From our observations, almost all pitching scenes fulfill the condition  $C_V$ , but not  $C_A$ . Thus, unlike in [15], we declare the  $i$ th frame as a pitching scene if the following expression yields TRUE:

$$C_{\text{pitch}} = C_V \cdot (C_L + C_A + C_P). \quad (12)$$

Here,  $+$  and  $\cdot$  indicate the boolean OR and AND operations, respectively. In [15], we assumed that all pitching scenes satisfy  $C_A$  and are used as an AND condition in  $C_{\text{pitch}}$ . Using the proposed logic, we successfully detect the pitching shots with an 80.6% accuracy for the baseball games under consideration.

## 4.2 Audio processing

### 4.2.1 Speech/non-speech classification

We use an unsupervised non-stationarity measure-based speech /non-speech classification scheme as presented in [15]. The approach is based on a long-term standard deviation calculated on the Mel-filter bank energies (MFBE) of the audio frames, which is sensitive to non-stationarity, and found to be efficient at distinguishing audio environments. If  $m_{ij}$  indicates the Mel-filter bank energy of the  $j$ th audio-frame and  $i$ th filter-bank, we use 40 filter banks (i.e.,  $i = 1, \dots, 40$ ) where each audio frame is 25-ms long with a 10-ms overlap, then signal non-stationarity is estimated by computing the standard deviation of MFBE over a longer time-period termed as segments. Here,  $\sigma_{kj}$  will be

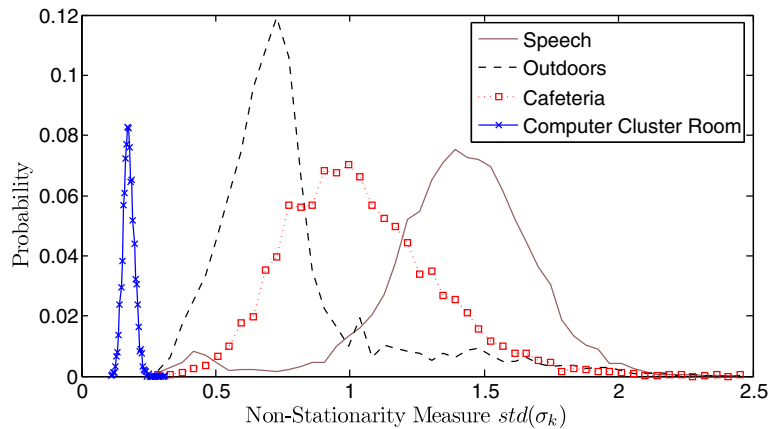
the  $k$ th standard deviation for the  $j$ th Mel-filter which is denoted by:

$$\sigma_{kj} = \sqrt{\frac{1}{N_s} \sum_{i=(k-1)N_s+1}^{kN_s} \left( m_{ij} - \frac{1}{N_s} \sum_{i=(k-1)N_s+1}^{kN_s} m_{ij} \right)^2}, \quad (13)$$

where  $N_s$  is the number of frames in the time period. In this study,  $N_s = 20$  is used (e.g., 200 ms total duration window). Next, we form a vector,  $\sigma_{\mathbf{k}} = (\sigma_{k1}, \sigma_{k2}, \dots, \sigma_{kM})$ , ( $M = 40$ ). The standard deviation of the components  $\sigma_{\mathbf{k}}$  given by  $std(\sigma_{\mathbf{k}})$  is found to be efficient at distinguishing audio environments. Using this measure, speech/background classification was performed on the game audio using a two mixture GMM as discussed in [15]. Figure 13 shows the probability distribution of  $std(\sigma_{\mathbf{k}})$  for speech and compares it to different acoustic environments such as outdoors (walking between different locations on campus), cafeteria (fully occupied during lunch hours), and computer cluster room (room consists of multiple servers). The distributions reveal that speech and background are easily separable within the feature space. Thus, a simple unsupervised segmentation algorithm that uses the proposed non-stationarity measure is presented. In the implementation,  $std(\sigma_{\mathbf{k}})$  is computed for each segment of the game video and a two-mixture GMM is trained using the non-stationarity measure utilizing the expectation-maximization (EM) algorithm. The underlying intuition exploited here is that one Gaussian would learn speech, while the other is expected to learn the background distribution characteristics. This learning is exploited by computing the posterior of each mixture component for every feature  $P_g(k)$  as

$$P_g(k) = \frac{1}{\sqrt{2\pi}\sigma_g} \exp \left[ -\frac{(std(\sigma_{\mathbf{k}}) - \mu_g)^2}{2\sigma_g^2} \right], \quad (14)$$

where  $\mu_g$  and  $\sigma_g^2$  are the mean and variance of the  $g$ th Gaussian ( $g = 1, 2$ ). Using the  $P_g(k)$  values, each segment is assigned to the more likely Gaussian (i.e., the one with the higher posterior probability). Since the non-stationarity of speech is typically higher than that of the background, the Gaussian with the larger mean is assumed to be speech. Using this intuition, speech and background Gaussians within the GMM are identified and every  $k$ th segment is assigned to either speech or background. Furthermore, Viterbi smoothing is used to smooth the above decisions using a high self-transition probability (0.9 ~ 0.99). This segmentation is used for our speech excitement level detection scheme. Using the above technique on audio data from the six baseball games (about 15 h of audio), an overall accuracy of 80.1% is obtained with a low miss rate of 2.6% (miss is speech



**Figure 13** Feature distribution for various acoustic environments.

detected as background) and false alarm rate of 17.3% (false alarm here represents the background detected as speech).

#### 4.2.2 Excitement measurement in speech

Using the features discussed in section 3, a GMM-based classifier is designed to detect high and moderate excitement levels from the game audio. Details of the evaluation of this scheme is presented later in section 5.1 within other evaluations from this study. This section describes how the excitement classification output is used in construction of the feature parameter for highlights extraction.

To estimate the required  $G_i(t)$  functions for high and moderate/low excitement in speech, we use the same principle used for slow motion feature. First, we form the function  $\Phi_4(t)$ , such that  $\Phi_4(t) = 1$  if the high excitement class was detected at time instant  $t$ , and 0, otherwise. Similarly, we form  $\Phi_5(t)$  for the detected moderate/low excitement class. Time domain examples of these functions are shown in Figure 13d,e. Next, the corresponding  $G_i(t)$  functions are computed following Equation 8. The only difference here is that the function  $\Phi_5(t)$  is inverted before filtering, following the fact that low excitement in speech is inversely proportional to the viewer arousal.

#### 4.2.3 Audio energy measure

To compute the audio energy, a fixed audio segment size of 267 samples is chosen to be equivalent to our video frame rate of 29.97 frames/section. For each frame  $t$ , the audio energy  $\Phi_6(t)$  is extracted and later filtered using Equation 8 to obtain  $G_6(t)$ . Finally, the averaged value of  $G_6(t)$  in the  $k$ th segment is used to compute the segmental audio energy features  $x_i(k)$ .

#### 4.3 Feature fusion and highlights generation

In order to generate the highlights time curve  $H_M(t)$  [10], the functions  $G_i(t)$  are filtered using a weighting function,

$$w(t) = \frac{1}{2} \left( 1 + \operatorname{erf} \left( \frac{\tilde{G}_M(t) - d}{\sigma} \right) \right) \quad (15)$$

to obtain,

$$G'_i(t) = G_i(t)w(t), \text{ with } i = 1, \dots, D \quad (16)$$

where  $D = 6$  is the number of features, and the choice of  $d = 40$  and  $\sigma = 5$  follows [10], and

$$\tilde{G}_M(t) = \min_i [\operatorname{sort}_i^M (\{G_i(t) | i = 1 \dots D\})]. \quad (17)$$

Here,  $\operatorname{sort}_i^M$  provides the top  $M$  values of  $G_i(t)$  at time location  $t$ . We use  $M = 3$  here. Finally, the highlights time curve is generated as,

$$H_M(t) = \left[ \frac{\max_t(a(t))}{\max_t(\tilde{a}(t))} \right] \tilde{a}(t), \quad (18)$$

where

$$a(t) = \sum_i \eta_i G'_i(t),$$

$$\tilde{a}(t) = K_t(l, \beta) * a(t), \text{ and}$$

$$\sum_i \eta_i = 1.$$

To estimate excitability in segment  $k$ , we use the averaged values of  $H_M(t)$  in that segment to obtain  $\bar{H}_M(k)$ .

For the proposed excitability measure, first, the multimodal feature vector  $\mathbf{x}(k) = (x_1(k), x_2(k) \dots x_D(k))$  is computed for each segment  $k$ . All features are normalized to zero mean with unit variance before further processing. Next, the histogram of the  $x_i(k)$  values across all segments of the video is used to estimate the PDFs  $f_{X_i}(\lambda)$ . The excitability measure  $\zeta(k)$  from each segment can then be

computed using Equation 3. Note that in this implementation,  $\zeta(k)$  measure is calculated without a need of any prior knowledge about the features  $x_i(k)$ . In order to generate a highlight video, segmentation is performed using the detected pitching shot locations. The proposed measure is then used to rank order and combine the video segments according to the user-defined overall highlights duration. Sample highlights generated using the proposed technique can be found in <http://sites.google.com/site/baseballhighlights/>.

#### 4.4 Generation of excitement time curve

In many applications, a continuous estimate of the excitement level is desired. The proposed technique can also be used to generate such a curve. In this case, the method is directly applied on  $G_i(t)$  functions instead of the segmental feature parameters  $x_i(k)$ . The pitching shot detection is no longer required for this analysis. A sample of such a highlights curve extracted from a baseball video segment of about 20 min is shown in Figure 14. The excitement time curve  $H_M(k)$  extracted following [10] is also shown for comparison. We note that the ground truth/real user excitement curve was not estimated for comparison. The plot in Figure 14 is simply a demonstration of the proposed method's ability to generate a continuous curve as in [10], as well as evaluating the excitement of a video segment.

#### 4.5 Real-time implementation

In real-time highlight generation, important highlight events (e.g., a home run or a goal, etc.) are detected and played back immediately or sent to the users through online media. These methods usually work on explicitly detecting the important events using game dependent cues. Since the proposed scheme functions on low-level features and estimates excitability using a probabilistic measure, real-time event detection and broadcast is not feasible in the current framework. However, the method can be extended to operate in real-time by processing video segments in chunks and updating the highlights periodically. Though on-line highlight generation is not

the focus of this study, we describe here how this can be achieved using the proposed framework.

In this scenario, the excitability measure  $\zeta(k)$  requires to be estimated in real-time from the past  $N$  segments as

$$\zeta(k) = -\log_2 \prod_{i=1}^D \int_{x_i(k)}^{\infty} \hat{f}_{X_i}(\lambda) d\lambda, \quad (19)$$

where  $\hat{f}_{X_i}$  indicates the PDF estimate of the  $i$ th feature at segment  $k$  obtained from the past  $M$  segments. The PDFs may also be estimated from all the video segments starting from the beginning of the game up to segment  $k$ . For finding a highlight event within the last  $M$  segments, a predefined threshold  $\tau$  on  $\zeta(k)$  can be used. This threshold can be estimated from the  $\zeta(k)$  values from the past  $M$  segments. This means that if the excitability measure found in the current segment  $k$  is significantly high compared to what was observed in the past  $M$  segments, segment  $k$  is used as highlight. This approach is summarized in Algorithm 1. It should be noted that, after the game is over, the highlights generated form the full video may not contain all the local 'short-time' highlight segments that were detected in real-time.

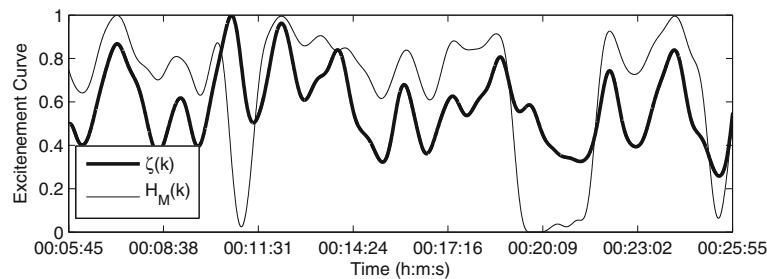
#### Algorithm 1 Real-time highlight generation using the proposed framework

---

```

Load the first  $M$  video segments as buffer;
Estimate feature PDFs;
for  $k = 1 \rightarrow M$  do
    | Compute  $\zeta(k)$  using (3);
end
for  $k = M + 1 \rightarrow last\_segment$  do
    | Re-estimate PDFs using feature segments:  $\{1 \dots k\}$ ;
    | Compute  $\zeta(k)$  using (19);
    | Set threshold  $\tau$ ;
    | if  $\zeta(k) > \tau$  then
        | | Select  $k$  as a highlight segment;
        | | Broadcast segment or notify user;
    | end
end
    
```

---



**Figure 14** Excitement time curve generated from a baseball game segment as in [10] ( $H_M(k)$ ) and using proposed surprisal measure ( $\zeta(k)$ ).

## 5 Evaluation of proposed system

In this section, we first evaluate the speech production-based classifier in terms of the excitability measure for sports commentators speech. Next, we perform a subjective evaluation of the complete audio-video highlights extraction system combining the multi-modal features.

### 5.1 Speech production-based classifier

As discussed in section 3, fundamental frequency  $F_0$ , spectral center of gravity  $SCG$ , and the first three formants  $F_{1-3}$  were found to provide good correlation with the perceptual excitement labels and hence are used in this section as input parameters for automatic excitement-level assessment.

The speech-based classifier utilized in the currently proposed highlights generation scheme replicates the one presented in our preliminary study [15]. Two GMMs are trained to represent 'moderate' excitement (referring to excitement levels 1 and 2) and 'high' excitement (levels 3 and 4). The 32-mixture GMMs are trained on samples extracted from 4 baseball games, where the subjective excitement-level labels serve as target transcriptions of the data, and evaluated on 2 remaining games representing the open test set. The binary classification results are evaluated by means of equal error rate (EER). In a round robin experiment, the classifier utilizing features normalized by the game level statistics provides EER of 21.4 to 22.4 %, see [15] for details.

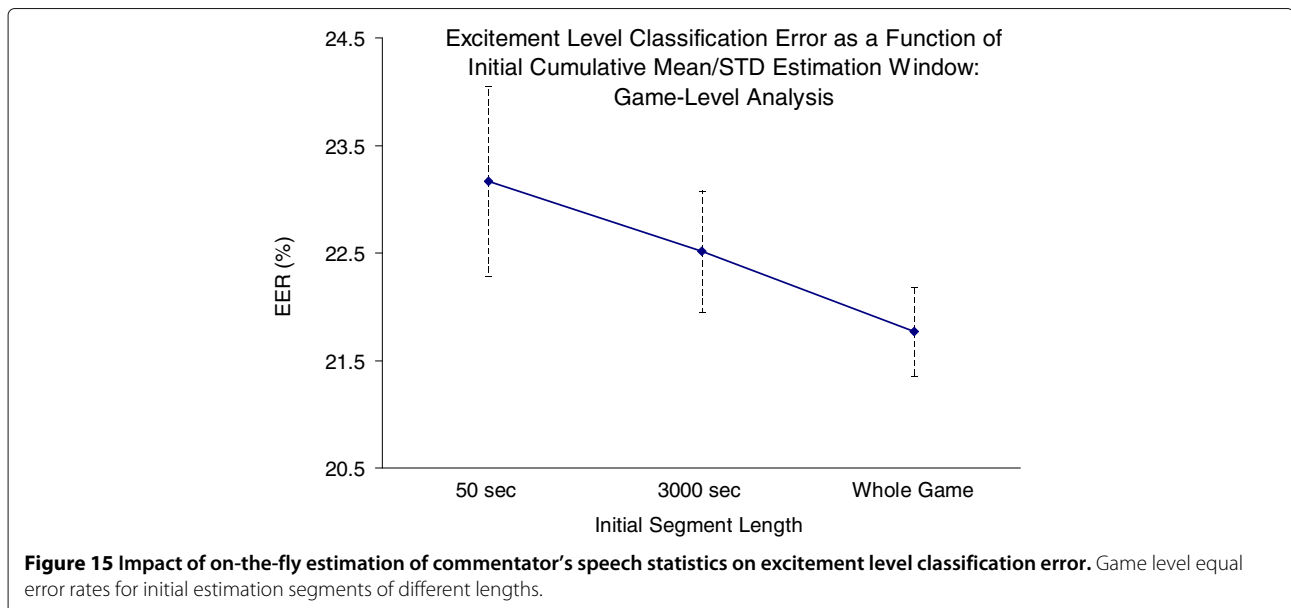
The feature normalization utilized in the previous paragraph assumes that the whole game is available to estimate the global statistics of the commentator's speech (see section 3). Some applications may require generation of highlight previews while the game is still in progress. In

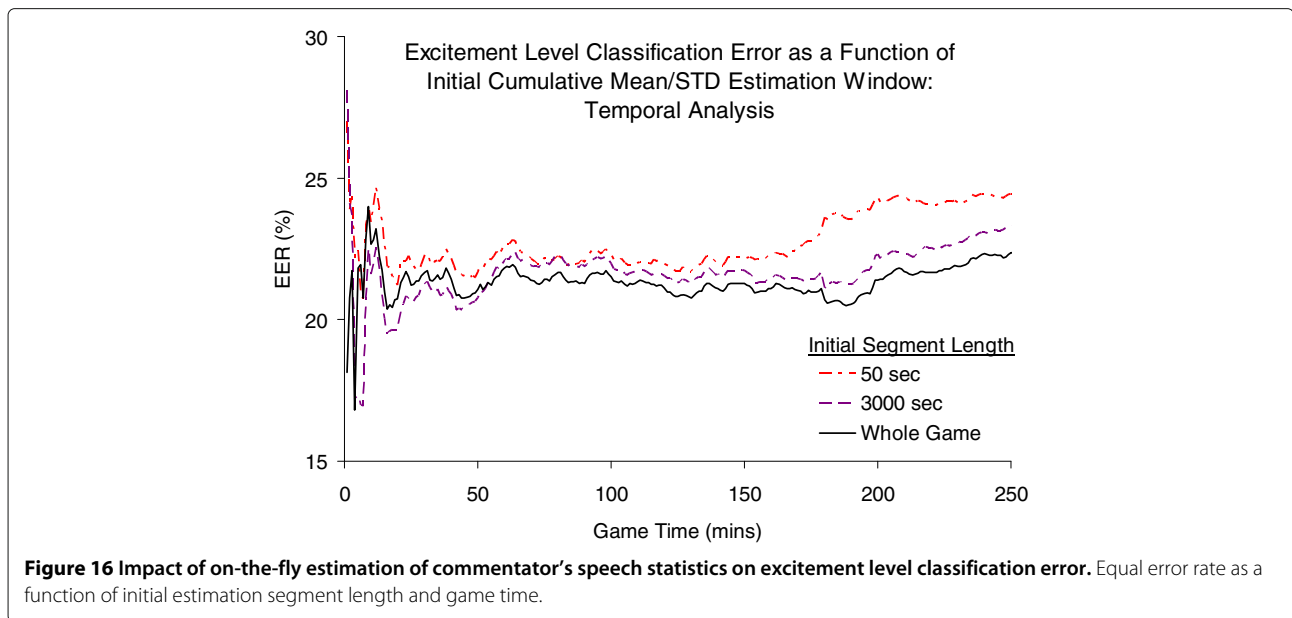
such a case, the statistics of commentator's speech features may be estimated on-the-fly. One possible approach is to estimate the feature means and standard deviations from the starting segment of the game and cumulatively update their values with every newly observed speech island. It is expected that the cumulative statistic estimates will converge over time towards the global game statistics.

To evaluate how the on-the-fly estimation might impact the accuracy of the excitement level classification, we implement a simple scheme where the statistics estimated from the initial segment of the game are continuously updated based on the newly observed speech islands using a weighted mean. The currently available estimates are used to normalize features extracted from the current voiced island. Clearly, there would be an option to go back in time and re-normalize all past segments with the newest estimates; however, we do not utilize this option in our current study.

The speech-based excitement level classification results for the on-the-fly approach are summarized in Figures 15 and 16. The first figure compares average game EERs for classifiers utilizing on-the-fly feature normalization (initial estimation segments of lengths 50 and 3,000 s) and a classifier employing game level statistics ('Whole Game'). The EERs are averaged across three round robin experiments and the vertical bars represent their standard deviations. It can be seen that there is a benefit from using game level statistics but the performance drop when switching to on-the-fly processing with an initial segment of 50 s is quite affordable (approximately 1.4 % absolute EER increase).

Figure 16 details the averaged round robin EERs as functions of the elapsed game time. A 1-min-long window was





applied to smooth the temporal EERs. It can be seen that the classifier utilizing game level statistics performs more accurately through a major part of the game; however, besides the first 4 min of the game, where the EER difference for both 50- and 3,000-s setups versus the game level setup drops by nearly 9 % EER, the EER difference for the global and on-the-fly systems does not exceed more than approximately 3 %, which seems quite affordable.

## 5.2 Subjective evaluation of the highlight generation system

In order to evaluate the effectiveness of the proposed measure of excitability, we conducted an independent subjective evaluation involving five viewers familiar with the game of baseball. We note that subjective evaluation of video summarization is quite challenging since it requires each subject to watch the entire game (typically approximately 3-h long), memorize all exciting moments and then select the most exciting parts to assess if the algorithm properly selected the highlights. In addition, it is natural to have a strong disagreement between viewers in deciding which scene should be selected/rejected. To make the evaluation feasible, we selected 100 pitching scenes (semantic segments) of about 15 s in duration from a baseball game. The subjects were asked to watch the videos and rank the excitability of the scene on a scale from 0 to 30. The rubrics used are boring (0 to 10), moderately exciting (10 to 20), and very exciting (20 to 30). Since the dynamic range of the ranking scores may vary across subjects, the scores obtained from each subject are normalized to zero mean and unit variance, and then averaged to obtain the final subjective measure. Note

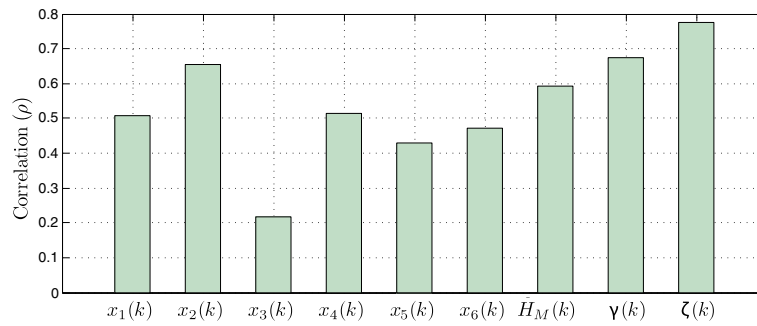
that the normalization here does not affect the ranking order of the segments decided by the subject (i.e., order from lowest ranking to highest ranking scenes will be preserved and also segments eligible to form highlights remain the same). Next, we extract excitement estimation in the selected 100 video segments obtained from (i)  $\bar{H}_M(k)$  computed using [10], (ii)  $\gamma(k)$  computed from a simple addition and normalization of the feature parameters  $x_i(k)$  [15], and (iii) the proposed measure  $\zeta(k)$ . The measures are evaluated by comparing their correlation and ranking agreement with the subjective measure.

### 5.2.1 Correlation between individual features and subjective scores

At first, we analyze how each feature parameter used in this study is correlated to the subjective user excitability. For data points,  $y_i$  and  $z_i$ ,  $i = \{1, \dots, N\}$ , the correlation coefficient  $\rho$  is calculated as

$$\rho = \frac{\text{cov}(y, z)}{\sigma_y \sigma_z} = \frac{\sum_{i=1}^N (y_i - \bar{y})(z_i - \bar{z})}{\sqrt{\sum_{i=1}^N (y_i - \bar{y})^2 \sum_{i=1}^N (z_i - \bar{z})^2}}. \quad (20)$$

This equation is used for calculating the correlation between the feature coefficients  $x_i(k)$  and the subjective scores ( $\zeta(k)$ ,  $\gamma(k)$ , or  $\bar{H}_M(k)$ ). The correlation of the individual features with the subjective scores are summarized in Figure 17 and Table 2. Here, we observe that all of the selected feature parameters  $x_i(k)$  are correlated with the subjective scores with positive values (i.e., they are proportional to excitability). Also, in each correlation value in Table 2, the probability of the null hypothesis ( $p$  value) is  $p < 0.05$  (i.e., the correlation is significant).



**Figure 17** Correlation coefficients of different feature parameters/objective excitement measures and subjective evaluation scores. Excitement measures are computed by fusing the feature parameters using equal weights.

Looking closely at Figure 17, we conclude that for our particular case, the most effective feature parameter is the average camera motion, with the highest individual correlation coefficient ( $\rho = 0.6531$ ). This is an indication that camera movement is more prominent during exciting moments of a baseball game, which is probably not the case for tennis, where the camera is more or less stationary. Excitement level detected in commentator speech is another good indication of excitability ( $\rho = 0.5146$ ). This is more or less true for any sports. The slow motion feature is also highly correlated with excitability ( $\rho = 0.5066$ ).

Observing the individual contribution of the features towards the excitability measure, we now like to analyze the effect of different feature weights  $\eta_i$  in (3). We use three different weight vectors  $\eta = [\eta_1 \dots \eta_D]$ :

1. Equal weights:  $\eta_i = 1/D = 1/6$ . Weight vector:  
 $\mathbf{w}_0 = [1 \ 1 \ 1 \ 1 \ 1 \ 1] / 6$
2. Higher weights to features  $x_1(k)$  &  $x_2(k)$ :  
 $\mathbf{w}_1 = [2 \ 2 \ 1 \ 1 \ 1 \ 1] / 8$
3. Higher weights to features  $x_2(k)$  &  $x_4(k)$ :  
 $\mathbf{w}_2 = [1 \ 2 \ 1 \ 2 \ 1 \ 1] / 8$

These weight settings are motivated by the higher correlation these individual features demonstrate (slow motion, high audio excitement and camera motion activity). In practical applications, we will not be able to estimate the

**Table 2** Comparison of correlation coefficient ( $\rho$ ) between different feature parameters and subjective evaluation scores

Segmental feature parameter	Correlation ( $\rho$ )	$p$ value
Slow motion duration, $x_1(k)$	0.5066	$7.559 \times 10^{-8}$
Average motion activity, $x_2(k)$	0.6531	$1.779 \times 10^{-13}$
Average scene cut density, $x_3(k)$	0.2171	$3.01 \times 10^{-2}$
Duration of speech in <i>high</i> class, $x_4(k)$	0.5146	$4.317 \times 10^{-8}$
(-ve) Duration of speech in <i>low</i> class, $x_5(k)$	0.4300	$7.992 \times 10^{-6}$
Audio energy, $x_6(k)$	0.4702	$7.952 \times 10^{-7}$

correlation values since the subjective scores will not be available. However, using a development dataset, these weights can be trained and later applied during evaluation.

### 5.2.2 Correlation between subjective and objective scores

When equal weights are used for the parameters  $\eta_i$  in (3), the proposed excitability measure  $\zeta(k)$  is seen to have the highest correlation coefficient of 0.77473, which outperforms the other two excitability measures  $\gamma(k)$  and  $\hat{H}_M(k)$ , yielding correlations of 0.67579 and 0.59287, respectively. This indicates the effectiveness of the proposed measure in fusing multiple feature parameters. These values are also shown in Figure 17, indicating that the subjective scores are better correlated with the proposed excitability measure, as opposed to the individual features.

Similar to the previous section, we compute these correlation coefficients for the alternative feature weights  $\mathbf{w}_1$  and  $\mathbf{w}_2$ . The results are summarized in Table 3, columns 6 and 7. With the weight setting of  $\mathbf{w}_1$ , the proposed method provides a correlation value of 0.80145, which is again superior compared to the other methods shown. The corresponding  $p$  values here again verify that the correlation values are significant ( $p < 0.05$ ).

### 5.2.3 Comparison between subjective and objective highlight rank

To evaluate objective rank ordering of the segments, we choose  $N$  top subjectively ranked video segments as a ground truth for highlights. We use the following values:  $N = \{10, 20, 30, 40\}$ . The excitement ranking measures  $\hat{H}_M(k)$ ,  $\gamma(k)$ , and  $\zeta(k)$  are then evaluated with respect to accuracy, true-positive rate (TPR) and false positive rate (FPR). For equal feature weights, the results are summarized in Figure 18. Here, again, we observe that the proposed measure provides superior performance compared to [10] and [15] for most values of  $N$ . For the alternative feature weights  $\mathbf{w}_1$  and  $\mathbf{w}_2$ , the TPR values for different  $N$ s are summarized in Table 3, columns 2 to 5. In each of these cases, we find that the proposed method provides superior performance (in italics).

**Table 3 Performance evaluation by comparing top  $N$  highlight segments selected using subjective and objective methods**

Method	%TPR in top $N$ segments				Correlation ( $\rho$ )	$p$ value
	10	20	30	40		
Using equal feature weights						
$\bar{H}_M(k)$ [10]	40.00	55.00	66.67	67.50	0.59287	$8.096 \times 10^{-11}$
$\gamma(k)$ [15]	40.00	55.00	66.67	75.00	0.67579	$1.224 \times 10^{-14}$
$\zeta(k)$ (Proposed)	50.00	70.00	73.33	77.50	0.77473	$3.183 \times 10^{-21}$
Using feature weights: $\eta = \mathbf{w}_1$						
$\bar{H}_M(k)$ [10]	50.00	55.00	63.33	67.50	0.60627	$2.319 \times 10^{-11}$
$\gamma(k)$ [15]	50.00	45.00	66.67	77.50	0.70183	$4.150 \times 10^{-16}$
$\zeta(k)$ (Proposed)	60.00	70.00	83.33	82.50	0.80145	$1.319 \times 10^{-23}$
Using feature weights: $\eta = \mathbf{w}_2$						
$\bar{H}_M(k)$ [10]	30.00	55.00	63.33	67.50	0.60197	$3.486 \times 10^{-11}$
$\gamma(k)$ [15]	40.00	45.00	63.33	75.00	0.68468	$4.011 \times 10^{-15}$
$\zeta(k)$ (Proposed)	40.00	60.00	70.00	77.50	0.77747	$1.880 \times 10^{-21}$

Correlation coefficient ( $\rho$ ) and corresponding  $p$  values are also shown.

As an example, for  $N = 20$ , the proposed technique selects 70% of the top 20 subjectively ranked segments with an FPR of 7.5%, while for [10], the TPR and FPR values are 55% and 11.2%, respectively. With proper design of the feature weights, the highlight selection performance can improve further as shown in Table 2.

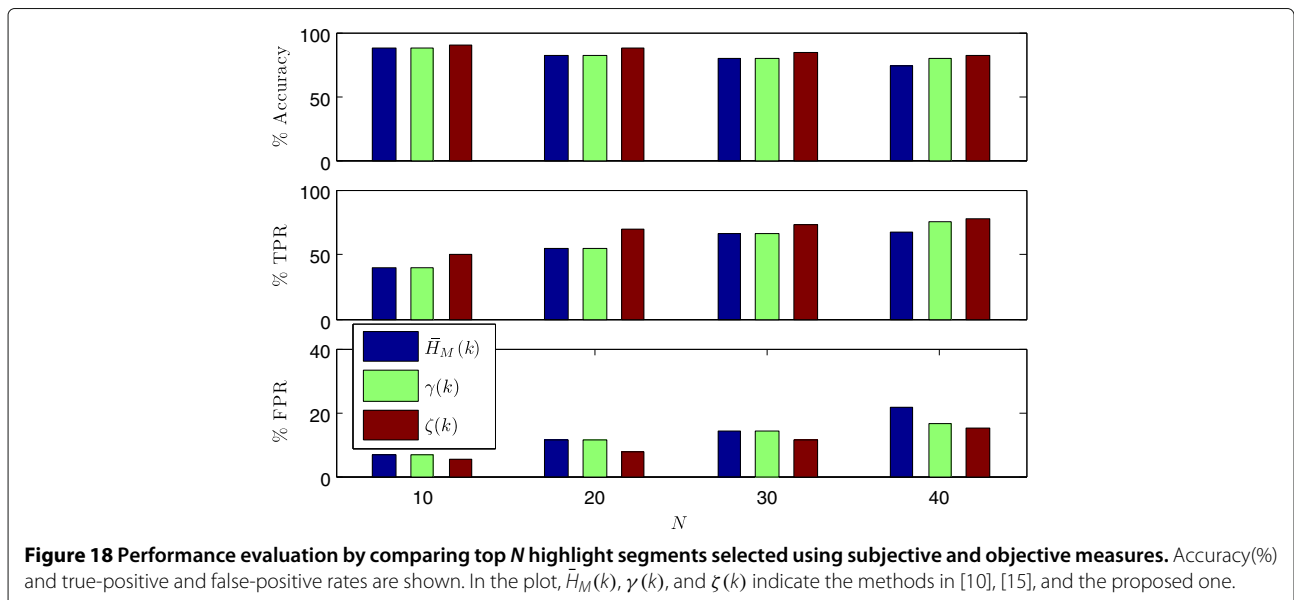
## 6 Conclusion

A generic video highlights generation scheme based on an information theoretic measure of user excitability was presented. The scheme utilizes audio excitement and low-level video features. Based on the analysis of the sports commentator's speech, production parameters most correlated with the perceptual excitability were selected to

form an acoustic feature vector for excitement assessment. Along with this, audio energy, slow motion duration, camera motion, and scene-cut density were used as features in the highlights extraction system. The proposed integrated excitability measure was shown to effectively combine the multi-modal features in video segments and found to be highly correlated with a perceptual assessment of excitability. The proposed scheme was also shown to outperform state-of-the-art generic excitability ranking methods.

## Endnote

<sup>a</sup>Outcomes of this study were presented in short forms in our preliminary studies [15,16].



**Figure 18 Performance evaluation by comparing top  $N$  highlight segments selected using subjective and objective measures.** Accuracy(%) and true-positive and false-positive rates are shown. In the plot,  $\bar{H}_M(k)$ ,  $\gamma(k)$ , and  $\zeta(k)$  indicate the methods in [10], [15], and the proposed one.



### Competing interests

The authors declare that they have no competing interests.

### Example Processed Output Content:

The following websites provide sample processed output time domain highlight compressed video content using the CRSS-UTDallas proposed system developed in this study.

<http://sites.google.com/site/baseballhighlights/>

<http://crss.utdallas.edu/demos/>

### Acknowledgements

This project was funded by the AFRL under contract FA8750-12-1-0188 (Approved for public release, distribution unlimited), and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J.H.L. Hansen.

Received: 8 November 2012 Accepted: 14 October 2013

Published: 18 November 2013

### References

1. H Pan, P Van Beek, M Sezan, Detection of slow-motion replay segments in sports video for highlights generation, in *Proc. IEEE ICASSP* 7–11 May 2001
2. M Delakis, G Gravier, P Gros, Audiovisual integration with Segment Models for tennis video parsing. *Comput. Vis. Image Underst.* **111**(2), 142–154 (2008)
3. M Fleischman, B Roy, D Roy, Temporal feature induction for Baseball highlight classification, in *Proceedings of the 15th Int. Conf. on Multimedia*, Augsburg, Germany, 24–29 Sept. (ACM, New York, 2007), pp. 333–336
4. Z Xiong, R Radhakrishnan, A Divakaran, Generation of sports highlights using motion activity in combination with a common audio feature extraction framework, in *Proc. IEEE ICIP*, Barcelona, Catalonia, 14–17 Sept., (2003)
5. M Kolekar, S Sengupta, Semantic concept mining in cricket videos for automated highlight generation. *Multimedia Tools and Appl.* **47**(3), 545–579 (2010)
6. D Tjondronegoro, X Tao, J Sasongko, C Lau, Multi-modal summarization of key events and top players in sports tournament videos, in *2011 IEEE Workshop on Applications of Computer Vision (WACV)*, Kona, HI, 5–7 January (IEEE, Piscataway, 2011), pp. 471–478
7. C Cheng, C Hsu, Fusion of audio and motion information on HMM-based highlight extraction for baseball games. *IEEE Trans. Multimedia* **8**(3), 585–599 (2006)
8. C Lien, C Chiang, C Lee, Scene-based event detection for baseball videos. *J. of Visual Comm. and Image Representation.* **18**, 1–14 (2007)
9. J Assfalg, M Bertini, AD Bimbo, W Nunziati, P Pala, Soccer highlights detection and recognition using HMMs, in *Proc. ICME '02*, Lausanne, Switzerland, 26–29 Aug. 2002 *Volume 1* (IEEE, Piscataway, 2002), pp. 825–828
10. A Hanjalic, Adaptive extraction of highlights from a sport video based on excitement modeling. *IEEE Trans. Multimedia* **7**(6), 1114–1122 (2005)
11. Z Xiong, R Radhakrishnan, A Divakaran, T Huang, Audio events detection based highlights extraction from baseball, golf and soccer games in a unified framework, in *Proc. IEEE ICASSP*, Hong Kong, China, 6–10 April 2003 *Volume 5* (IEEE Computer Society, Washington, DC, 2003), pp. 632–635
12. P Chang, M Han, Y Gong, Extract highlights from baseball game video with hidden Markov models, in *Proc. Image Process.*, Rochester, New York, 22–25, Sept. *Volume 1* (IEEE, Piscataway, 2002), pp. 609–612
13. Y Rui, A Gupta, A Acero, Automatically extracting highlights for TV baseball programs, in *Proceedings of the Eighth ACM International Conference on Multimedia*, Los Angeles, CA, October 30 – November 03 (ACM, New York, 2000), p. 115
14. Y Ma, L Lu, H Zhang, M Li, A user attention model for video summarization, in *Proceedings of the tenth ACM international conference on Multimedia*, Juan les Pins, France, 1–6 December, 2002 (ACM, New York, 2002), p. 542
15. H Bořil, A Sangwan, T Hasan, JHL Hansen, Automatic excitement-level detection for sports highlights generation, in *Proc. InterSpeech*, Makuhari, Chiba, Japan 26–30 September 2010, pp. 2202–2205
16. T Hasan, H Bořil, A Sangwan, JHL Hansen, A multi-modal highlight extraction scheme for sports videos using an information-theoretic excitability measure, in *IEEE ICASSP, Kyoto, Japan 25–30, March 2012*, pp. 2381–2384
17. TM Cover, JA Thomas, *Elements of Information Theory* (Wiley-Interscience, New York, 1991)
18. JHL Hansen, Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition. *Speech Comm.* **20**(1–2), 151–173 (1996)
19. R Cowie, Douglas-E Cowie, N Tsapatsoulis, G Votsis, S Kollias, W Fellenz, JG Taylor, Emotion recognition in human-computer interaction. *IEEE Signal Process. Mag.* **18**, 32–80 (2001)
20. H Bořil, T Kleinschmidt, P Boyraz, JHL Hansen, Impact of cognitive load and frustration on drivers' speech. *The J. Acoust. Soc. Am.* **127**(3), 1996–1996 (2010)
21. CM Lee, SS Narayanan, Toward detecting emotions in spoken dialogs. *IEEE Trans. on Speech & Audio Process.* **13**(2), 293–303 (2005)
22. K Sjolander, J Beskow, Wave Surfer—an open source speech tool, in *Proc. of ICSLP'00*, Beijing, China, 16–20 Oct. 2000 *Volume 4*
23. R Schulman, Dynamic and perceptual constraints of loud speech. *The J. Acoust. Soc. Am.* **78**(S1), S37–S37 (1985)
24. P Gramming, S Sundberg, S Ternström, W Perkins, Relationship between changes in voice pitch and loudness. *STL-QPSR* **28**, 39–55 (1987)
25. Y Lu, M Cooke, Speech production modifications produced by competing talkers, babble, and stationary noise. *The J. Acoust. Soc. Am.* **124**(5), 3261–3275 (2008)
26. D Pisoni, R Bernacki, H Nusbaum, M Yuchtman, Some acoustic-phonetic correlates of speech produced in noise, in *Proc. of ICASSP*, Tampa, Florida, 26–29 March, *Volume 10* (IEEE, Piscataway, 1985), pp. 1581–1584
27. RD Kent, C Read, San Diego, *The Acoustic Analysis of Speech* (Whurr Publishers, 1992)
28. Z Bond, T Moore, A note on Loud and Lombard speech, in *Proc. of ICSLP'90*, Kobe, Japan, 18–22 November 1990
29. H Bořil, Robust speech recognition: analysis and equalization of Lombard effect in Czech Corpora, PhD thesis (Czech Technical University in Prague, Czech Republic, 2008). <http://www.utdallas.edu/~hxb076000>
30. JC Junqua, The Lombard reflex and its role on human listeners and automatic speech recognizers. *The J. Acoust. Soc. Am.* **93**, 510–524 (1993)
31. H Wakita, Normalization of vowels by vocal-tract length and its application to vowel identification. *IEEE Trans. Acoust. Speech and Signal Processing.* **25**(2), 183–192 (1977)
32. A Oppenheim, R Schaffer, *Discrete-Time Signal Processing* (Prentice Hall, Upper Saddle River, NJ, 1999)
33. C Liu, Q Huang, S Jiang, L Xing, Q Ye, W Gao, A framework for flexible summarization of racquet sports video using multiple modalities. *Comput. Vis. Image Underst.* **113**, 415–424 (2009)
34. H Liu, WJ Zhang, J Cai, A fast block-matching algorithm based on variable shape search. *J. Zhejiang University - Science A* **7**, 2006. [10.1631/jzus.2006.A0194]
35. B Truong, C Dorai, S Venkatesh, New enhancements to cut, fade, and dissolve detection processes in video segmentation, in *Proc. of the 8th ACM Inter. Conf. on Multimedia*, Los Angeles, CA, October 30 – November 03, 200 (ACM, New York, 2000), p. 227
36. W Chu, J Wu, Explicit semantic events detection and development of realistic applications for broadcasting baseball videos. *Multimedia Tools and Appl.* **38**, 27–50 (2008)
37. R Van Den Boomgaard, R Van Balen, Methods for fast morphological image transforms using bitmapped binary images. *Graphical Models and Image Process.* **54**(3), 252–258 (1992)

doi:10.1186/1687-6180-2013-173

Cite this article as: Hasan et al.: Multi-modal highlight generation for sports videos using an information-theoretic excitability measure. *EURASIP Journal on Advances in Signal Processing* 2013 **2013**:173.