Modelling Sports Highlights Using a Time Series Clustering Framework and Model Interpretation

Regunathan Radhakrishnan, Isao Otsuka, Ziyou Xiong, Ajay Divakaran

TR2005-077   December 2005

Abstract

In our past work on sports highlights extraction, we have shown the utility of detecting audience reaction using an audio classification framework. The audio classes in the framework were chosen based on intuition. In this paper, we present a systematic way of identifying the key audio classes for sports highlights extraciton using a time series clustering framework. We treat the low-level audio features as a time series and model the highlight segments as ünusualëvents in a background of an üsualp̈rocess. The set of audio classes to characterize the sports domain is then identified by analyzing the consistent patterns in each of the clusters output from the time series clustering framework. The distribution of features from the training data so obtained for each of the key audio classes, is parameterized by a Minimum Description Length Gaussian Mixture Model (MDL-GMM). We also interpret the meaning of each of the mixture components of the MDL-GMM for the key audio class (the ḧighlightc̈lass) that is correlated with highlight moments. Our results show that the ḧighlightc̈lass is a mixture of audieance cheering and commentatorś excited speech. Furthermore, we show that the precision-recall performance for highlights extraction based on this ḧighlightc̈lass is better than that of our previous approach which uses only audience cheering as the key highlight class.

SPIE Storage and Retrieval Methods and Applications for Multimedia

Modelling Sports Highlights using a Time Series Clustering Framework & Model Interpretation

Regunathan Radhakrishnan†, Isao Otsuka‡, Ziyou Xiong†and Ajay Divakaran†
†Mitsubishi Electric Research Laboratory,
Cambridge, MA 02139
‡Mitsubishi Electric Corporation, Japan
E-mail: †{regu,zxiong,ajayd}@merl.com
‡otsuka@img.kyo.melco.co.jp

[illegible]

In our past work on sports highlights extraction, we have shown the utility of detecting audience reaction using an audio classification framework.[6] The audio classes in the framework were chosen based on intuition. In this paper, we present a systematic way of identifying the key audio classes for sports highlights extraction using a time series clustering framework. We treat the low-level audio features as a time series and model the highlight segments as "unusual" events in a background of an "usual" process. The set of audio classes to characterize the sports domain is then identified by analyzing the consistent patterns in each of the clusters output from the time series clustering framework. The distribution of features from the training data so obtained for each of the key audio classes, is parameterized by a Minimum Description Length Gaussian Mixture Model (MDL-GMM). We also interpret the meaning of each of the mixture components of the MDL-GMM for the key audio class (the "highlight" class) that is correlated with highlight moments. Our results show that the "highlight" class is a mixture of audience cheering and commentator's excited speech. Furthermore, we show that the precision-recall performance for highlights extraction based on this "highlight" class is better than that of our previous approach which uses only audience cheering as the key highlight class.

[illegible]

Past work on sports highlights extraction using audio cues has mostly focussed on detecting specific events that are correlated with highlights. For baseball, Rui et al have detected the announcer's excited speech and bat-ball impact sound using a directional template matching based on audio signal only.[5] For golf, Hsu has used Mel Scale Frequency Cepstrum Coefficients as audio features and a multivariate Gaussian as a classifier to detect Golf club-ball impact.[4]

We proposed a unified audio classification framework for extracting sports highlights from different sports including soccer,golf and baseball.[6] The audio classes (applause, cheering, music, speech and speech with music) in the proposed framework were chosen to characterize different kinds of the sounds that were common to all of the sports. For instance, the first two classes were chosen to capture the audience reaction to interesting events in a variety of sports.

In this paper, we propose a systematic way to acquire domain knowledge to arrive at the audio classes that would characterize the different sounds in a given domain. We treat the low-level audio features as a time series and use the time series clustering framework proposed in[3] to identify a key audio class ("highlight" class). The time series clustering framework in[3] models highlights as "unusual" events in a background of a "usual" process. After we collect sufficient data for the "highlight" audio class using the clustering framework, we train a Minimum Description Length-Gaussian Mixture Model to parameterize the distribution of features of the audio class.

The rest of the paper is organized as follows. In section 2, we provide a brief introduction to the time series clustering framework. In section 3, we present the Minimum Description Length criterion to arrive at the number of mixture components in GMMs for modelling audio classes. In section 4, we present our experimental results on several baseball and soccer games and finally conclude in section 5.

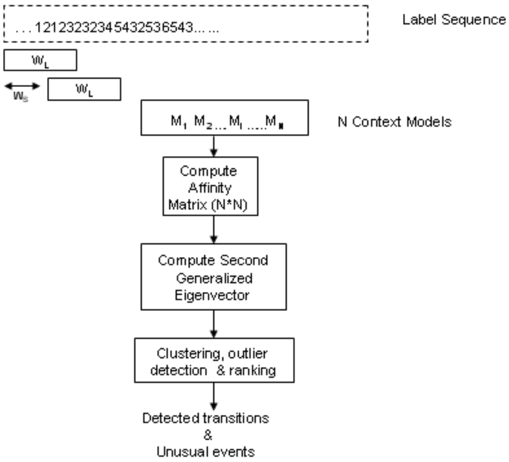
. . . 12123232345432536543... ...
Label Sequence
W_L
W_S
W_L
M_1 M_2... M_i ....M_N
N Context Models
Compute Affinity Matrix (N*N)
Compute Second Generalized Eigenvector
Clustering, outlier detection & ranking
Detected transitions & Unusual events

Figure 1. [illegible]

2 [illegible]

2 [illegible]

In this section, we present a brief description of the time series clustering framework proposed in.[3] The proposed framework is motivated by the observation that "interesting" events in multimedia happen sparsely in a background of usual or "uninteresting" events. Some examples of such events are:

- [illegible]: A burst of overwhelming audience reaction following a highlight event in a background of commentator's speech.
- [illegible]: A burst of laughter following a comical event in a background of dialogues.
- [illegible]: A burst of motion and screaming noise following a suspicious event in a silent or static background.

This motivates us to formulate the problem of mining for "interesting" events in multimedia as that of detecting outliers or "unusual" events by statistical modelling of a stationary background process in terms of low/mid-level audio-visual features. Note that the background process may be stationary only for small period of time and can change over time. This implies that background modelling has to be performed adaptively throughout the content. It also implies that it may be sufficient to deal with one background process at a time and detect outliers.

2 [illegible]

Given the problem of detecting times of occurrences of "unusual" events from a time series of observations from several processes of which there is one dominant background process, we propose the following time series clustering framework:

- Sample the input time series on a uniform grid. Let each time series sample (consisting of a sequence of observations) be referred to as context.
- Statistically model the time series observations within each context.
- Compute the affinity matrix for the whole time series using the context models and a commutative distance metric defined between two context models. The affinity matrix represents a graph where each node is a model and the weight on the edge connecting two nodes is $\exp(-\frac{d}{\sigma^2})$ where d is the defined distance metric and the σ parameter controls how quickly the similarity falls off.
- Use the normalized cut solution (the second generalized eigenvector of the computed affinity matrix) to identify distinct clusters & outlier context models.

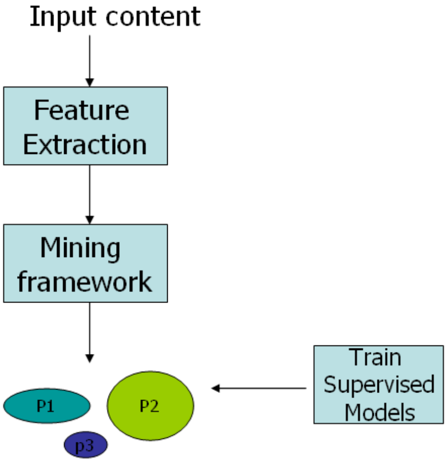

Figure 2. [illegible]

Figure 1 illustrates the above mining framework. In this framework, there are three key issues namely the statistical model for the context and the choice of the two parameters, W_L and W_S. The choice of the statistical model for the time series sample in a context would depend on the underlying background process. A simple unconditional PDF estimate would suffice in case of a memoryless background process. However, if the process has some memory, the chosen model would have to account for it. For instance, a Hidden Markov Model would provide a first order approximation.

The choice of the two parameters (W_L and W_S) would be determined by the confidence with which one wants to say that something is an "unusual" event.The size of the window W_L determines the reliability of the statistical model of a context.The size of the sliding factor, W_S, determines the resolution at which unusual event is detected. Please refer,[3] for our analysis on how W_L determines the confidence on the detected unusual event.

2 [illegible]

The proposed mining framework can be used for systematic acquisition of domain knowledge thereby making the choice of semantic classes to characterize the domain less ad-hoc. In this section, we describe one such exercise with the audio of sports domain. Given the audio stream of a sports video clip, we extract low-level spectral or cepstral features and treat them as a time series. Using the time series clustering framework, we can first obtain distinguishable sound classes for the chosen features. Then, by examining individual clusters one can identify consistent patterns in the data that correspond to the events of interest and build supervised statistical learning models.

Figure 3 shows an example of how such a framework can be used for the selection of semantic classes. In this figure, the second generalized eigenvector of the affinity matrix for a sports clip shows outliers at times of occurrences of applause segments. Furthermore, there are two distinct clusters for the segments corresponding to different speakers. These clusters are irrelevant as far as highlights extraction is concerned and hence can be grouped together under a single speech label by training a speech GMM using training data collected from different speakers. Such an analysis brings out the interaction between different clusters and enables the choice of the relevant semantic classes that can help detect the target concept as an outlier.

Once target semantic classes have been chosen, one can use GMMs to parameterize the distribution of features. In the following section, we present the theory behind MDL-GMMs which is one of the methods to arrive at the number of mixture components for GMMs.

3 [illegible]

The model parameters of MDL-GMMs are obtained by minimizing the Rissanen's objective function that trades-off between model complexity and goodness of fit of the model to the observed data. For audio classification

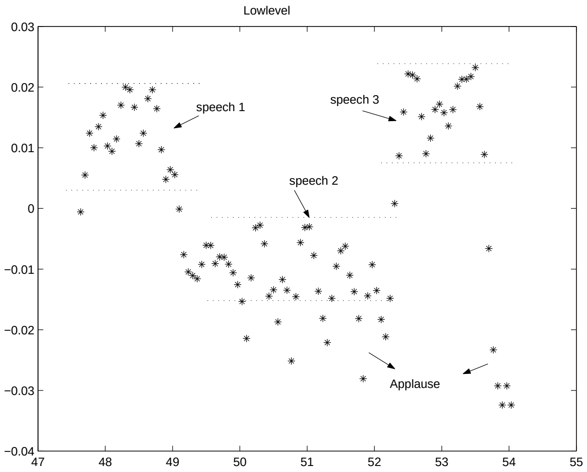

Figure 3. [illegible]

based highlights extraction, it has been shown in[7] that MDL-GMMs outperform GMMs with number of mixture components chosen arbitrarily.This motivates us to use MDL-GMMs for modelling the distribution of features of the "highlight" audio class as well.

The objective function for obtaining the optimal number of mixture components and model parameters can be derived as follows.[2] Let Y be an M dimensional random vector to be modelled using a Gaussian mixture distribution. Let K denote the number of Gaussian mixtures, and we use the notation π, μ, and R to denote the parameter sets $\{\pi_k\}_{k=1}^K$, $\{\mu_k\}_{k=1}^K$ and $\{R_k\}_{k=1}^K$ for mixture coefficients, means and variances. The complete set of parameters are then given by K and $\theta = (\pi, \mu, R)$. The log of the probability of the entire sequence $Y = \{Y_n\}_{n=1}^N$ is then given by

$$\log p_y(y|K,\theta) = \sum_{n=1}^{N} \log \left(\sum_{k=1}^{K} p_{y_n|x_n}(y_n|k,\theta)\pi_k \right) . \tag{1}$$

The objective is then to estimate the parameters K and $\theta \in \Omega^{(K)}$. The maximum likelihood (ML) estimate is given by

$$\hat{\theta}_{ML} = \arg \max_{\theta \in \Omega^{(K)}} \log p_y(y|K,\theta)$$

the estimate of K is based on the minimization of the expression

$$MDL(K,\theta) = -\log p_y(y|K,\theta) + \frac{1}{2} L \log(NM) , \tag{2}$$

where L is the number of continuously valued real numbers required to specify the parameter θ. In this application,

$$L = K \left(1 + M + \frac{(M+1)M}{2} \right) - 1 .$$

For details on parameter update rules, please refer.[7]

4. [illegible]

In this section, we present our results on some soccer games and baseball games recorded from Japanese as well as American broadcastings. From the accompanying audio track of the input sports video, low-level audio cepstral features were extracted for every frame of duration $8ms$. Each feature vector is a 13 dimensional feature

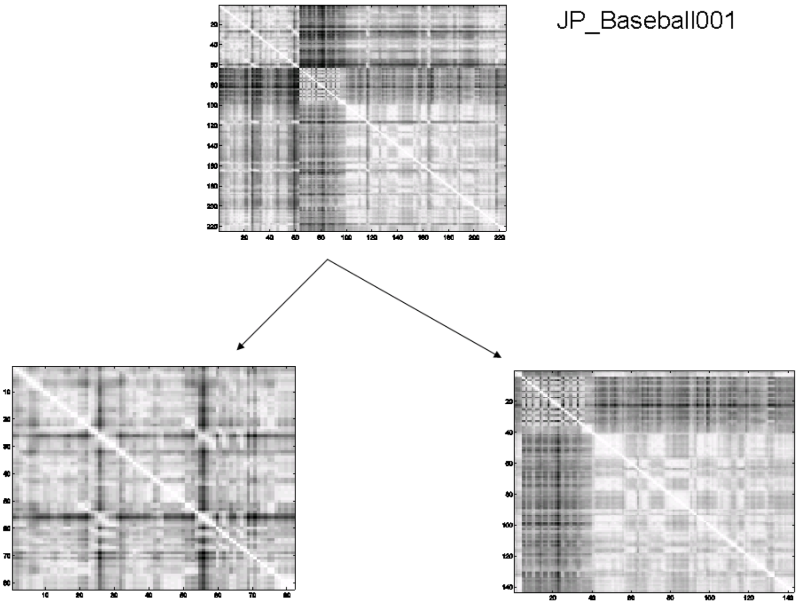

Figure 4. [illegible]

vector with 12 Mel Frequency Cepstral Coefficients (MFCC) and logarithm of energy of the frame. Then, we treat this sequence of observations as a time series input to our clustering framework.

The chosen size of W_L for mining this sequence of features was 8 seconds with a skip size (W_S) of 4 seconds. The chosen context model was a 2-component Gaussian Mixture Model. The input time series is then sampled every 4 seconds and a 2-component GMM is estimated from the observations within the 8 second window. Then, an affinity matrix is constructed by using the following distance metric between two context models, λ_1 and λ_2, with observation sequences O_1 and O_2 respectively.

$$\begin{aligned} D(\lambda_1, \lambda_2) = \frac{1}{W_L}(&\log P(O_1|\lambda_1) + \log P(O_2|\lambda_2) \\ &- \log P(O_1|\lambda_2) - \log P(O_2|\lambda_1)) \end{aligned}$$

The first two terms in the distance metric measure the likelihood of training data given the estimated models. The last two cross terms measure the likelihood of observing O_2 under λ_1 and vice versa. If the two models are different, one would expect the cross terms to be much smaller than the first two terms.

Then, we perform a hierarchical clustering by using normalized cut on this graph in the following way. We first partition the constructed graph into two individual clusters. Then, we construct the affinity matrix for the two identified clusters from the parent affinity matrix by simply picking the corresponding rows from the parent affinity matrix.

Figure 4 shows the affinity matrices in a hierarchical representation for 15 min from a Japanese Baseball game. For each node in the hierarchical representation, we can again use spectral clustering to come up with two clusters as shown in figure 5

We observed that the outliers in one of the partitions correspond to highlight moments. They were observed to be a mixture of audience cheering and commentator's excited speech suggesting that highlight moments are times when both the audience and commentator are excited. Our earlier highlights extraction framework assumed at the outset, that a cheering or an applause audio class would be sufficient to get to the "interesting" parts of

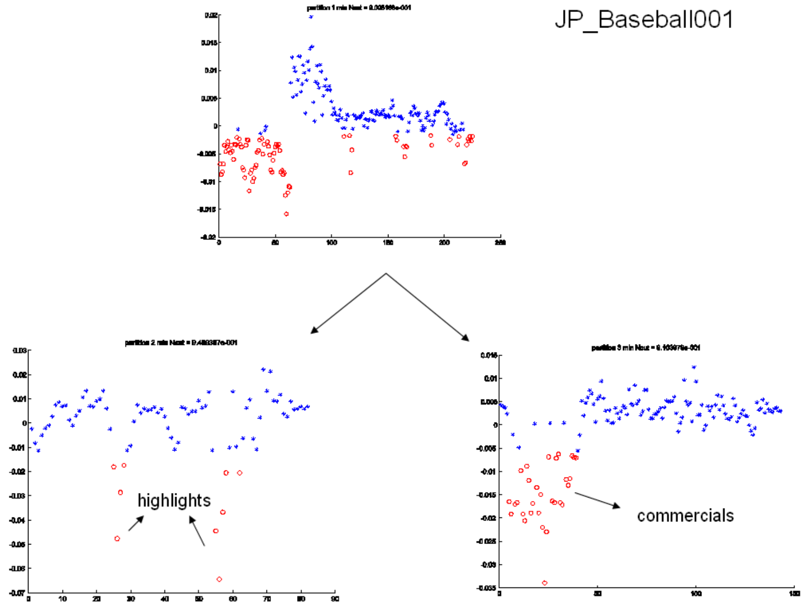

Figure 5. [illegible]

the video. However, this observation tells us that when there is only cheering from the audience it is less likely to be really interesting without the commentator also getting excited.

This motivated us to repeat the mining experiment with a few more games to collect more training data for this "highlight" audio class. We consistently got the same pattern from the other soccer and baseball games from Japanese and American sports content also as shown in figures **?** -7.

This procedure gave us sufficient amount of training data for the "highlight" audio class. We trained a Gaussian Mixture model using the minimum description length principle to model the distribution of low-level cepstral features. We use the percentage of this "highlight" audio class labels to rank the time segments in the input sports video as shown in figure 8. In the same figure, we have also shown the ranking for the time segments using percentage of cheering and applause labels. By choosing the same threshold on these two ranked streams, we select the time segments that have a rank greater than the threshold and verify manually if they were really highlight moments. We get a better Precision-Recall performance using the "highlight" class than simply using cheering and applause as shown in figure 9.

Now that we have a "highlight" class that gives us a better Precision-Recall performance, we proceed to interpret the meaning of the MDL-GMM of this class by inferring what each component is modelling for the given the training data set. The MDL solution for the number of components in the GMM for the "highlight" audio class data set was 4. Given an input feature vector, y_n and a K component GMM with θ as learned parameters, we can calculate the probability that a mixture component, k, generated y_n by using Bayes' rule as given below:

$$p(k/y_n,\theta) = \frac{p(y_n/k,\theta)\pi_k}{\sum_{k=1}^{K} p_{y_n}(y_n|k,\theta)\pi_k}$$

Then, we can assign y_n to the mixture component for which the posterior probability ($p(k/y_n,\theta)$) is maximum. If we append all the audio frames corresponding to each of the mixture components, we can interpret the semantic meaning of every component.[1] We performed mixture component inferencing for the "highlight" audio class

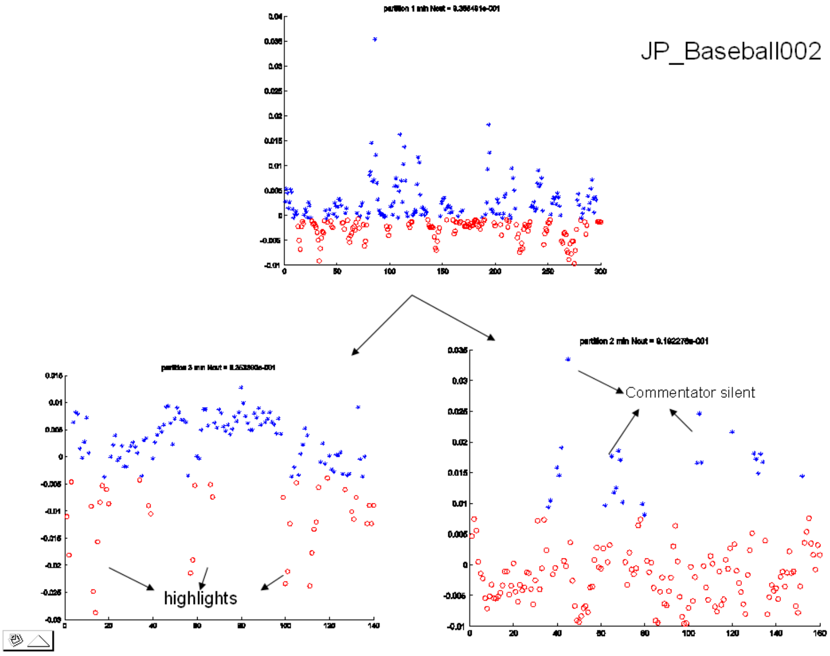

Figure 6. [illegible]

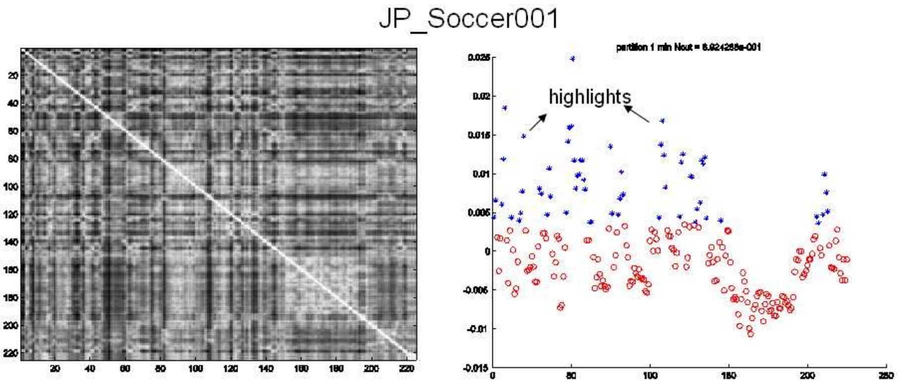

Figure 7. A [illegible]

using the MDL-GMM. We find that one of the components predominantly models the excited speech of the commentator and another component models the cheering of the audience.

5 [illegible]

In this paper, we have presented a systematic way to collect training data for modelling highlight moments in sports audio. We model the highlight moments as "unusual events" in a "usual" background. We treat the low-level audio features as a time series and perform a hierarchical clustering operation which detects highlight moments as outliers. We collected training data for the "highlight" audio class by performing the mining operation on several games. We used a MDL-GMM to model the distribution of features of the highlight audio class. We interpreted the semantic meaning of each of the mixture components of the learnt model to find out that the "highlight" class models the excited speech of the commentator as well as the cheering of the audience jointly. Using this "highlight" audio class, we have reduced the false alarm rate of our previous highlight extraction scheme which was based on cheering of the audience alone. The use of the new "highlight" class

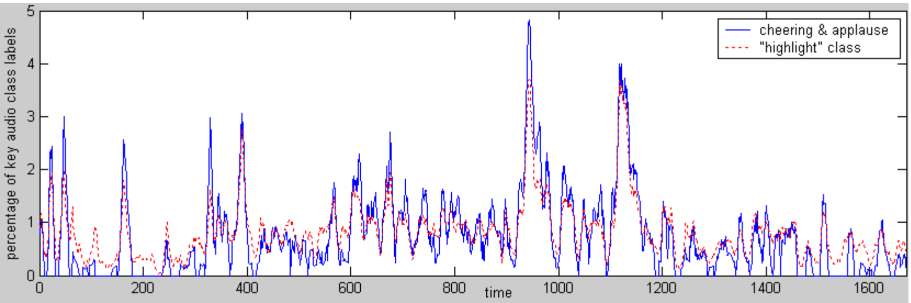

Figure 8. [illegible]

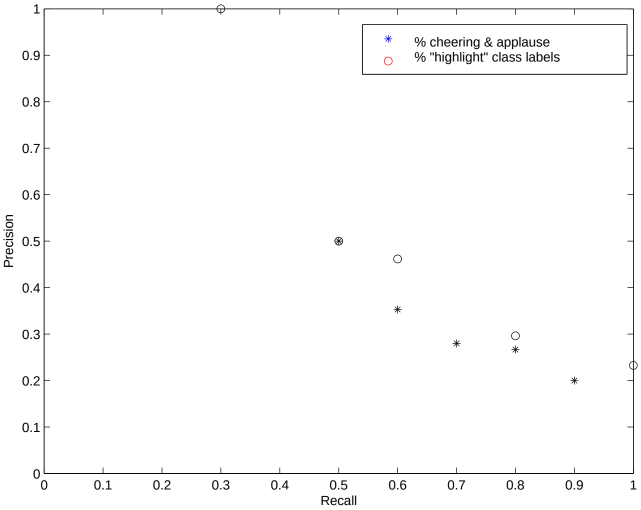

Figure 9. [illegible]

for ranking the time segments has been shown to be effective for summarizing baseball and soccer games from different content providers in America and Japan.

[illegible]

1. A. Poritz. Linear predictive hidden markov models and the speech signal. *Proc. of ICASSP* (1982).
2. Bouman, C. A. CLUSTER: An unsupervised algorithm for modeling gaussian mixtures. *http://www.ece.purdue.edu/∼bouman*. School of Electrical Engineering, Purdue University.
3. R.Radhakrishnan, A.Divakaran and Z.Xiong. A technical report on time series clustering based framework for multimedia mining and summarization. *http://www.merl.com/papers/TR2004-046/, MERL Technical Report* (2004).
4. W.Hsu. Speech audio project report. *Class Project Report,2000* (www.ee.columbia.edu/ winston).
5. Y.Rui,A.Gupta and A.Acero. Automatically extracting highlights for tv baseball programs. *Proc. of ICME* (2000).
6. Z. Xiong, R. Radhakrishnan, A. Divakaran and T.S. Huang. Audio-based highlights extraction from baseball, golf and soccer games in a unified framework. *Proc. of ICASSP* (2003).
7. Z.Xiong,R.Radhakrishnan, A.Divakaran and T.S.Huang. Sports highlights extraction using the minimum description length criterion in selecting gmm structures. *ICME* (2004).