# Excited Commentator Speech Detection with Unsupervised Model Adaptation for Soccer Highlight Extraction

Yi Sun[1], Zhijian Ou[1], Wei Hu[2], Yimin Zhang[2]
[1] *Department of Electronic Engineering, Tsinghua University, Beijing 100084, China*
[2] *Intel China Research Center, Beijing 100080, China*
*Email: ozj@tsinghua.edu.cn, wei.hu@intel.com*

## Abstract

*Soccer highlight detection is an active research topic in recent years. In this paper, we present our effort to detect an important audio keyword - excited commentator speech, which contributes to a state-of-the-art soccer highlight extraction system. We propose an approach of using statistical classifier based on Gaussian mixture models (GMMs) with unsupervised model adaptation. The excited speech and normal speech are modeled as two GMMs, and are updated to compensate for the acoustic mismatch between training and test data via Maximum a posteriori (MAP) adaptation, starting from the pre-trained GMMs. The adaptation is operated in an unsupervised mode, since the correct classification of the test data is not known, and a first pass of detection using old GMMs is performed to produce hypothesized classification results. Experimental results demonstrate the effectiveness of the proposed approach. Based on the excited speech detection alone, we can recall 87% of the goal events.*

## 1. Introduction

The extensive amount of multimedia information available necessitates the development of content analysis and summarization techniques to facilitate access and browsing. Broadcast soccer video is such a popular program that attracts a lot of researches [1-10]. Through automated highlight extraction, the consumers can retrieve interesting events (e.g. goal, shoot, free kick) quickly from the long videos and save time. However, highlight extraction is still a challenging problem due to the semantic gap between low-level features and high-level semantic events. To address such problem, a widely used approach is to employ a three-level framework [1, 2, 4], which consists of low-level audiovisual feature extraction, middle-level semantic keyword detection and high-level event extraction. Examples of middle-level keywords include view type, playing field position for the visual modality, and excited commentator speech, whistle for the audio modality. It is clear that the accuracy of middle-level keyword detection is crucial to the overall performance of such three-level system. In this paper, we present our effort to detect an important audio keyword - excited commentator speech, which contributes to a state-of-the-art soccer highlight extraction system [9, 10].

It is found that excited commentator speech is one of the most reliable indications of highlight events in soccer videos [3, 4, 5]. An excited commentary almost always corresponds to an interesting moment of the game. Typically, there are two types of methods for excited commentator speech detection: learning-based and rule-based. Learning-based methods [3, 4] either employ Bayes decision theory based on generative models, e.g. Gaussian mixture model (GMM), or directly train discriminative classifiers, e.g. support vector machine (SVM). Rule-based methods use simple acoustic features (energy, band energy, pitch, etc) and ad-hoc threshold [5]. Generally speaking, learning-based methods are more theoretically sound and give better performance than rule-based methods, while rule-based methods are faster and easy to implement.

However, learning-based methods often suffer from environment mismatch between training and testing. The learned classifier may perform much worse on unseen acoustic conditions in training. The challenge is how to cope with the great varieties of audio signals in soccer video. The audio track consists of commentator speech, mixed with audience noises, music noises, and automatic gain control changing audio levels. Another problem is that previous works assume long speech excitement segment and ignore the short but significant excitement. For example, in soccer games, short excited speech could indicate events such as sudden shoots and severe fouls.

To address these problems, we propose an approach of using statistical classifier based on GMMs with unsupervised model adaptation for excited

commentator speech detection. The idea is to update the GMMs to compensate for the acoustic mismatch between training and test data, starting from the pre-trained GMMs. The adaptation is operated in an unsupervised mode, since the correct classification of the test data is not known, and a first pass of detection using old GMMs is used to produce hypothesized classification results. This detection and adaptation procedure could operate in several rounds, i.e. iteratively[1].

The rest of the paper is organized as follows. In section 2, the proposed approach is presented in detail. Experimental results are given in section 3, followed by the conclusion in section 4.

## 2. Excited Commentator Speech Detection

In this paper, we mainly consider excited commentator speech detection for soccer games. It is easy to apply this approach to other kinds of sports games without much change.

### 2.1. Excited and Normal Speech Modeling: GMMs

Note that there is observable difference between excited and normal speech segment (Fig 1). For excited speech, there are three main changes:

a) The pitch is much higher than normal,
b) The total energy is higher
c) The energy distribution is different from normal speech in that the energy ratio in higher frequency range (typically 2000-3000Hz) is much larger.

However, pitch or energy can only serve as a filter to discard audio segment other than speech excitement, they cannot be used as a reliable gauge for detecting speech excitement due to some reasons. First, pitch rising may be caused by some emotion other than excitement (e.g. surprise). Second, there are usually loud cheers and background noises in sports videos, which make the estimate of commentator speech energy very hard. So instead of using several separate features, we employ GMMs to model both the excited speech and the normal speech segment.

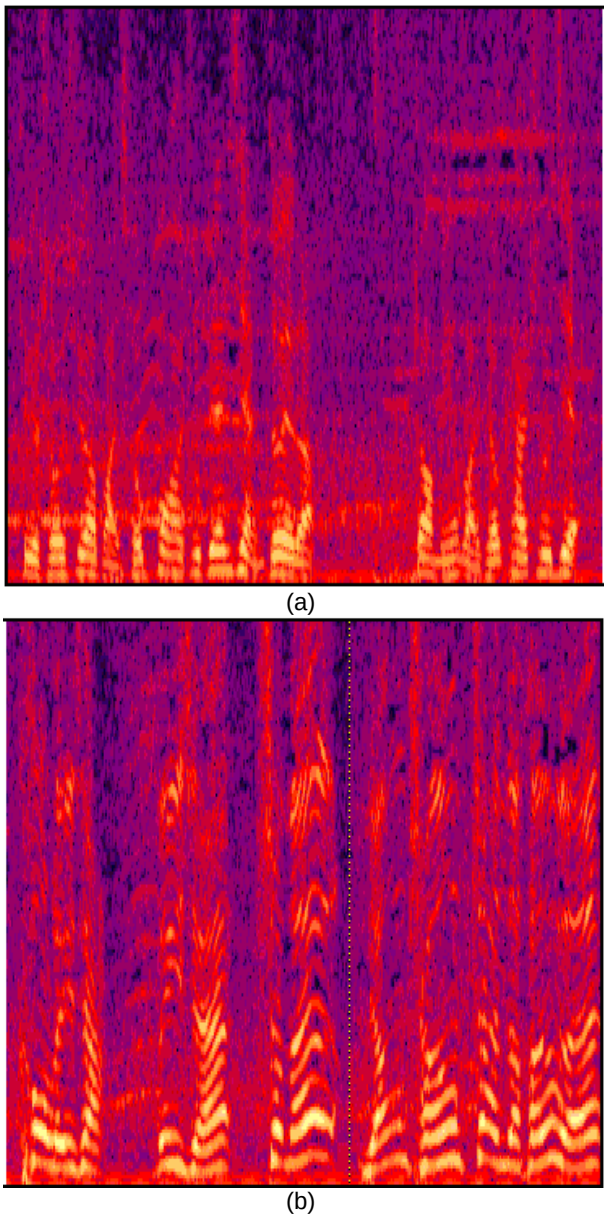


**Fig. 1. Spectrum of normal speech (a), and excited speech (b)**

For a D-dim feature vector $x$, the GMM density is define as:

$$P(x \mid \lambda) = \sum_{k=1}^{K} \omega_k N(x \mid \mu_k, \Sigma_k) \quad (1)$$

$$\lambda = \{\omega_k, \mu_k, \Sigma_k\}_{k=1}^{K} \quad (2)$$

Here $N(x|\mu,\Sigma)$ is a Gaussian distribution with mean $\mu$ and diagonal covariance matrix $\Sigma$. $K$ is the Gaussian mixture number and $\omega_k$ is the weight corresponding to the $k$-th mixture. Collectively, the parameters of the GMM are denoted as $\lambda$ as shown in (2).

[1] This multi-pass processing is not a problem, since highlight extraction can be done in an offline manner.

Given a collection of training vectors, maximum likelihood model parameters can be estimated using the iterative expectation–maximization (EM) algorithm [11].

## 2.2. MAP adaptation

Denote $\lambda_e$, $\lambda_n$ as the pre-trained GMMs for the excited speech and normal speech respectively. It is desirable that $\lambda_e$ and $\lambda_n$ well match the acoustic conditions in testing. However, this is rarely the case, especially for high-varied and noisy soccer audio.

To compensate for the acoustic mismatch between training and test data, the GMMs, $\lambda_e$ and $\lambda_n$, can be adapted to reflect the current test conditions. Since the correct classification of the test data is not known, we adopt unsupervised adaptation, which uses detection hypotheses to provide the adaptation supervision.

Given the supervision data $X$, the MAP estimate of model parameter $\lambda$, is

$$\begin{aligned}\lambda_{MAP} &= \arg\max_{\lambda} P(\lambda \mid X) \\ &= \arg\max_{\lambda} P(X \mid \lambda) P(\lambda)\end{aligned} \tag{3}$$

The formula for applying MAP estimation to GMM is well developed [12]. In our case, we adapt the Gaussian means and variances as in the following, while the mixture weights are kept fixed.

$$\hat{\mu}_k = (1-\alpha_1)\mu_k + \alpha_1 \frac{\sum_{t=1}^{T} P(k \mid x_t) x_t}{\sum_{t=1}^{T} P(k \mid x_t)} \tag{4}$$

$$\hat{\sigma}_k^2 = (1-\alpha_2)\left(\mu_k^2 + \sigma_k^2\right) + \alpha_2 \frac{\sum_{t=1}^{T} P(k \mid x_t) x_t^2}{\sum_{t=1}^{T} P(k \mid x_t)} - \hat{\mu}_k^2 \tag{5}$$

Here $T$ is the number of supervision features, $P(k|x_t)$ is the posterior probability of $k$-th Gaussian mixture calculated using the old GMM, $\alpha_1$, $\alpha_2$ is the adaptation factor tuned empirically.

## 2.3. Excited Speech Detection

Suppose we have trained two seed GMMs, $\lambda_e$ and $\lambda_n$ via EM algorithm, for excited speech and normal speech respectively, using some hand-annotated data. These two GMMs describe the average behavior of the normal speech and excited speech, and provide the prior information for the detection.

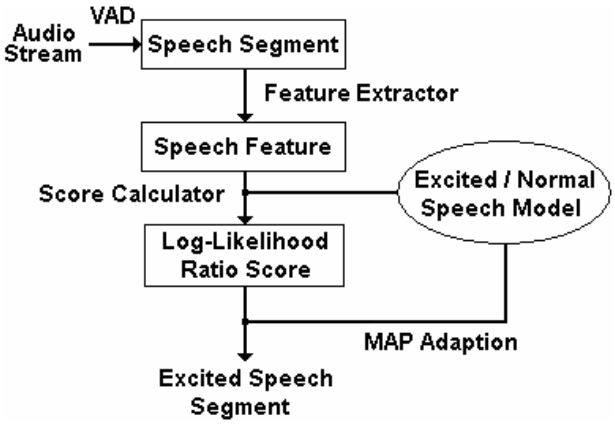


**Fig. 2. Flowchart of excited speech detection**

The detection process is shown in Fig. 2. It takes an audio stream as input and generates the excited speech segment as output. Since the seed models are trained with pitched speech segment, VAD (voice activity detection) is employed to filter out non-pitched segment. Next, features are extracted from the pitched speech segments and stored for later processing.

After feature extraction, an Unsupervised Model Adaptation is used to adjust the pre-trained GMMs. We first calculate the average log-likelihood ratio score for each pitched segment using their features, based on $\lambda_e$ and $\lambda_n$. Then, we do partitioning. According to the average log-likelihood ratio, we can partition all pitched speech segments into three groups:

a) Segments that is extremely close to excited speech (speech with very high score);
b) Segments that is extremely close to normal speech (speech with very low score);
c) Segments with average scores.

Next, we use features from most-likely excited speech segments (group a) and most-likely normal speech segments (group b) to update the respective GMMs via MAP adaptation, generating $\lambda_e^*$ and $\lambda_n^*$.

Finally, excited speech segments are chosen according to average log-likelihood ratio scores calculated from the updated GMMs $\lambda_e^*$ and $\lambda_n^*$.

## 2.4. Highlight Extraction Using Excited Commentator Speech Detection

There are efforts to infer the high-level highlight events from the generated middle-level keywords (including excited commentator speech) [9, 10]. In this paper, we use a simple post-processing to extract highlight, based on the excited speech detection result.

Specifically, we score each excited speech segment according to its duration and its log-likelihood score using (6), where $\tau$ is the duration of excited speech segment and *llr* is the log-likelihood ratio score.

$$score = \tau^{1/\beta} \cdot llr \quad (6)$$

The scoring function can be tuned to either bias towards short and extreme excited speech segment by increasing $\beta$ (such as shout) or favor long duration but only medium excited speech vice versa. Then, the chosen speech segments are merged according to their mutual distances and sorted by score. The first $N$ segments are extracted as highlight events.

## 3. Experimental results

We conduct two experiments to evaluate the performance of the excitement speech detection and the highlight extraction respectively.

We choose a 19-dim feature, which consists of a 14-dim MFCCs (Mel Frequency Cepstrum Coefficients) vector, a 4-dim pitch vector and an energy dimension. The MFCCs are known to give a good description of speech spectrum envelope. The 4-dim pitch vector includes the fundamental frequency (i.e. $f_0$), its first and second differentials, and a confidence value.

The seed GMM for excited speech is trained with hand-annotated data from 8 soccer half-matches. The GMM for normal speech is trained with data from the Mandarin Broadcast News corpus (Hub4-NE) [13], which is close to normal commentary speech but with less background noise. Each GMM has a mixture number of 128.

### 3.1. Excited Speech Detection Results

We use 6 half-matches (different from the training data) to test the accuracy of the proposed approach for excited speech detection. In each half-match, 15 or 20 most excited speech segments are detected as candidates. 62 out of 90 and 87 out of 120 are correct, giving an average accuracy of 68.9% and 72.5%. The results are shown in Table 1.

**Table 1. Accuracy for excited speech detection**

| No. of candidates per half-match | 15 | 20 |
|---|---|---|
| Positive in match 1 | 7 | 12 |
| Positive in match 2 | 10 | 14 |
| Positive in match 3 | 11 | 15 |
| Positive in match 4 | 11 | 16 |
| Positive in match 5 | 11 | 13 |
| Positive in match 6 | 12 | 17 |
| Overall accuracy (%) | 68.9 | 72.5 |

Decisions on whether a speech segment is correctly detected are made by several users. This result demonstrates that the overall performance of the proposed approach is acceptable, despite there is accuracy dropping in some cases (e.g. in match 1). Further inspection shows that this performance drop occurs due to strong background noises and nasal resonance in commentary speech. Since the test matches are randomly chosen and are even deliberately chosen from different broadcasters with different commentary styles, it is reasonable to infer that our proposed approach could suit a large amount of soccer games with satisfying result.

### 3.2. Highlight Extraction Results

To evaluate the performance of the highlight extraction algorithm, we use a dataset containing 13 half-matches with a total of 23 goals. For convenience and saving annotation time, we assume that only goals are labeled as highlight events and calculate the corresponding recall rate (ratio of the number of goals found compared to the total number of goals) based on this assumption. Note that such evaluation method may lead to performance underestimation, since goals are not the only highlight events containing excited speech segments. The result is listed in Table 2. In the test, we fixed the maximum number of candidates to 5 and 10, and result shows that a maximum of 87% goals are successfully recovered.

The influence of adding unsupervised adaptation (UA) is also demonstrated in Table 2. Through unsupervised adaptation, the recall rates improve by 4.4% and 9.5% for 10 and 5 candidates respectively.

**Table 2. Performance of highlight detection**

| No. of candidate per half match | 10 | | 5 | |
|---|---|---|---|---|
| Total goals | 23 | | | |
| Goals found (No UA/UA) | 19 | 20 | 14 | 16 |
| Recall (%) (No UA/UA) | 82.6 | 87.0 | 60.1 | 69.6 |

We also have to mention that as a stand-alone system, we limit the number of candidates to not more than 10. If we use such system as a module in a more advanced highlight extraction system and increase the maximum number of candidates, its performance would no doubt be better.

## 4. Conclusions

In this paper, we present our effort to detect an important audio keyword - excited commentator speech, which contributes to a state-of-the-art soccer highlight extraction system. We propose an approach of using statistical classifier based on Gaussian mixture models (GMMs) with unsupervised model adaptation. The excited speech and normal speech are modeled as two GMMs, and are updated to compensate for the acoustic mismatch between training and test data via Maximum

a posteriori (MAP) adaptation, starting from the pre-trained GMMs. The adaptation is operated in unsupervised mode, since the correct classification of the test data is not known, and a first pass of detection using old GMMs is used to produce hypothesized classification results. Experimental results demonstrate the effectiveness of the proposed method. Based on the excited speech detection alone, we can recall 87% of the goal events. It is worth pointing out that the proposed approach can be broadly applied to other kinds of sports videos.

## Acknowledgement

This work was supported by an Intel Corporation Grant.

## References

[1] Z. Xiong, R. Radhakrishnan, A. Divakaran, and T.S. Huang, "Effective and Efficient Sports Highlights Extraction Using the Minimum Description Length Criterion in Selecting GMM Structures," *ICME Conference*, 2004.

[2] Z. Xiong, R. Radhakrishnan, and A. Divakaran, "Generation of Sports Highlights Using Motion Activity in Combination with a Common Audio Feature Extraction Framework," *IEEE International Conference on Image Processing (ICIP)*, Vol. 1, 2003.

[3] Y. Rui, A. Gupta, and A. Acero, "Automatically extracting highlights for TV baseball programs," *ACM Multimedia Conference*, 105-115, 2000.

[4] F. Coldefy and P. Bouthemy, "Unsupervised soccer video abstraction based on pitch, dominant color and camera motion analysis," *ACM Multimedia Conference*, 2004.

[5] D. Tjondronegoro, Y.P. Chen, and B. Pham, "Sports Video Summarization using Highlights and Play-Breaks," *Proc. of the 5th ACM SIGMM international workshop on Multimedia information retrieval*, 2003.

[6] J. Wang, C. Xu, E. Cheng, K. Wan, and Q. Tian, "Automatic replay generation for soccer video broadcasting," *ACM Multimedia Conference*, 2004.

[7] A. Ekin, A.M. Tekalp, and R. Mehrotr, "Automatic soccer video analysis and summarization," *IEEE Trans. on Image processing*, 796-807, 2003.

[8] L. Xie, S.F. Chang, A. Divakaran, and H. Sun, "Structure analysis of soccer video with hidden Markov models," *Proc. ICASSP*, 4096-4099, 2002.

[9] T. Wang, J. Li, Q. Diao, W. Hu, and Y. Zhang, "Semantic Event Detection using Conditional Random Fields," *ICME Conference*, 2006.

[10] J. Li, T. Wang, W. Hu, M. Sun, and Y. Zhang, "Soccer Highlight Detection Using Two-dependence Bayesian Network," *ICME Conference*, 2006.

[11] A.P. Dempster, N.M. Laird, D. Rubin, "Maximum-likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, Series B, 39, 1977.

[12] J.L. Gauvain and C.H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Proc*., no.2, pp. 291-298, April 1994.

[13] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC98S73