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

Yi Sun, Zhijian Ou
Department of Electronic Engineering, Tsinghua University, Beijing, China

Wei Hu, Yimin Zhang
Intel China Research Center, Beijing, China

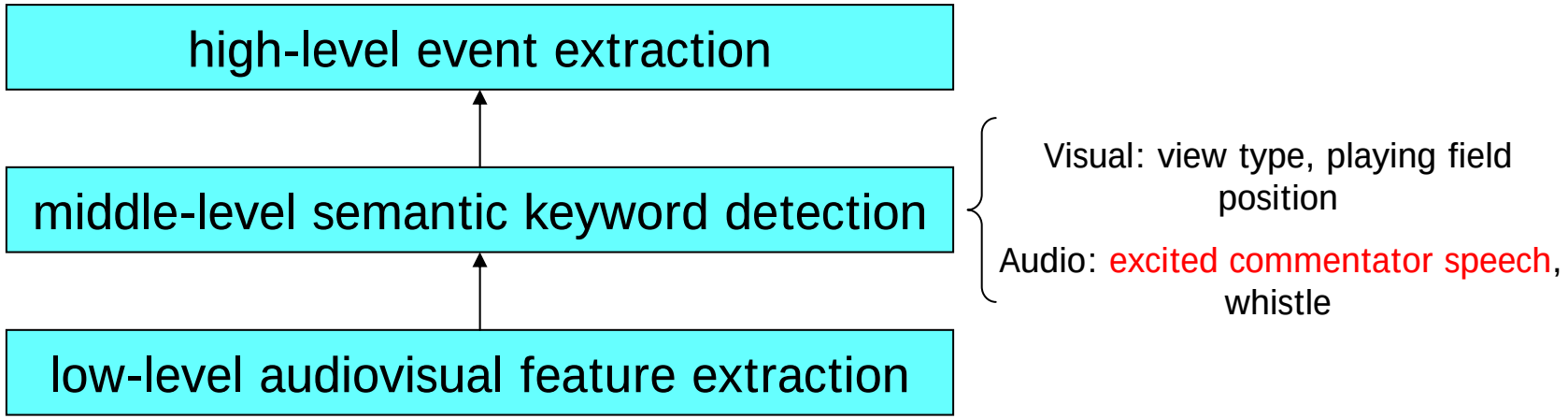
2011-11-24, ICALIP 2010, Shanghai

Content

- Introduction
- Excited Commentator Speech Detection for Soccer Highlight Extraction
  - Excited and Normal Speech Modeling: GMMs
  - Unsupervised MAP adaptation
- Experimental results
  - The proposed approach is simple yet effective

Introduction

- Task: automated soccer highlight extraction
  - Users can retrieve interesting events (e.g. goal, shoot, free kick) quickly from the long videos and save time.
- Difficult
  - Semantic gap between low-level features and high-level semantic events
- Three-level framework

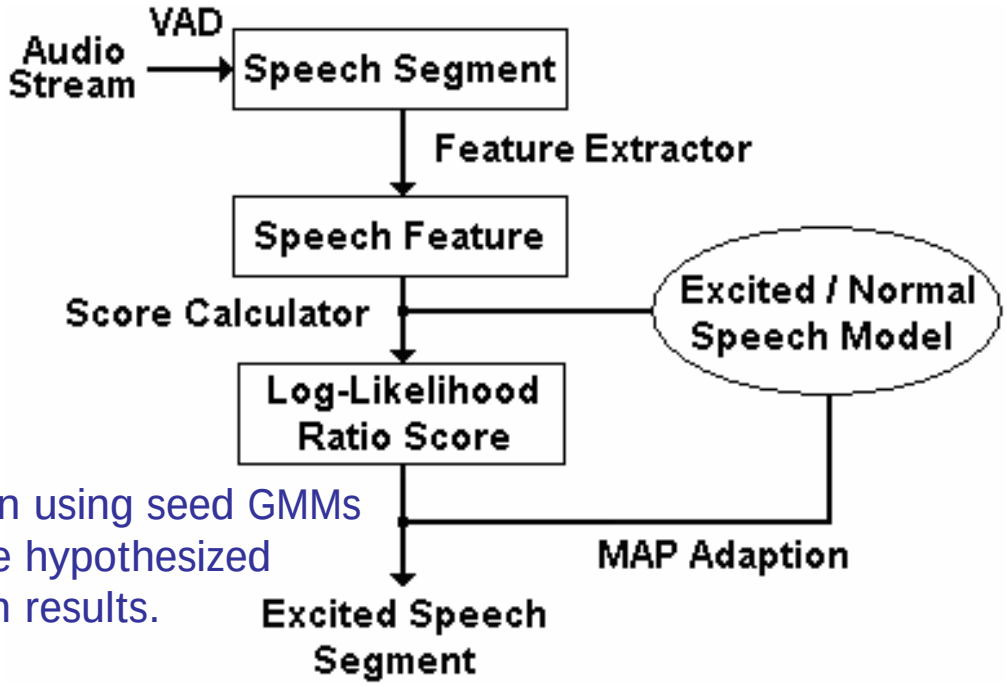


Excited speech detection

- Excited commentator speech
  - One of the most reliable indications of highlight events in soccer videos.
  - An excited commentary almost always corresponds to an interesting moment of the game.
- Existing problems with excited speech detection
  - The learned classifier may perform worse on unseen acoustic conditions in testing.
  - How to cope with the great varieties of audio signals in soccer video (mixed with commentator speech, audience noises, music noises, and automatic gain control changing audio levels.)

The proposed approach

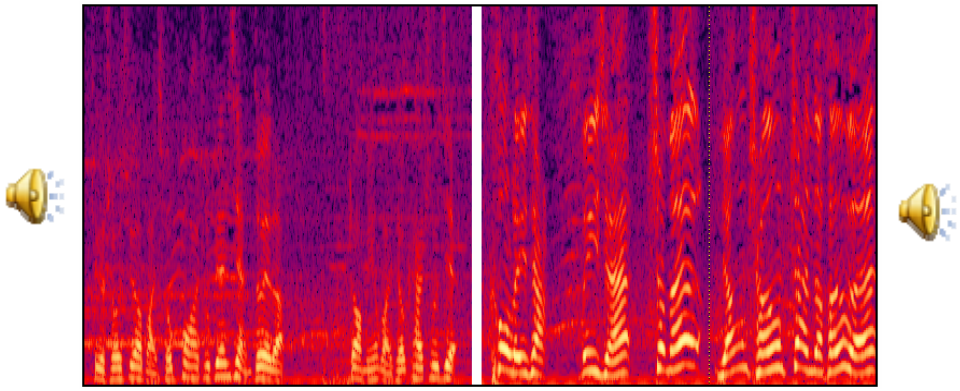
Propose an approach of using statistical classifier based on GMMs with unsupervised model adaptation



Two seed GMMs are trained offline from annotated audio

A first pass of detection using seed GMMs is used to produce hypothesized classification results.

Excited and Normal Speech Modeling: GMMs



Normal speech Excited speech

- Gaussian Mixture Model (GMM)

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

- 19-dim acoustic feature
  - 14-dim MFCCs (Mel Frequency Cepstrum Coefficients)
  - 4-dim pitch vector (f_0, Δ, $\Delta\Delta$, voicing degree)
  - energy

MAP adaptation

- Given the supervision data X, the MAP estimate of model parameter λ, is

$$\lambda_{MAP} = \arg\max_{\lambda} P(\lambda \mid X)$$
$$= \arg\max_{\lambda} P(X \mid \lambda) P(\lambda)$$

- MAP estimate formula for GMMs

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

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

Highlight Extraction

- From middle-level keywords (excited speech) to high-level highlight events
- Each excited speech segment is scored

$$score = \tau^{1/\beta} \cdot llr$$

  - τ is the duration of excited speech segment
  - llr is the log-likelihood ratio
  - β: a tuning paramter. Adjust β to tradeoff between short, extreme excited speech and long, but only medium excited speech

Experimental results

- Excited Speech Detection results
  - 8 half-matches for training seed GMMs
  - 6 half-matches for testing

No. of candidates per half-match	20
Positive in match 1	12
Positive in match 2	14
Positive in match 3	15
Positive in match 4	16
Positive in match 5	13
Positive in match 6	17
Overall accuracy	72.5%

Experimental results

- Highlight Extraction Results
  - Only goals are counted as highlight events

No. of candidate per half match	10	
Total goals	23	
Recall rate	87.0% (with adaptation)	82.6% (no adaptation)

Based on the excited speech detection alone, we can recall 87% of the goal events.

Conclusion

- Excited Commentator Speech Detection for Soccer Highlight Extraction
  - Excited and Normal Speech Modeling: GMMs
  - Unsupervised MAP adaptation
- Experimental results
  - The proposed approach is simple yet effective

Thank you !