SPEAKER SPOTTING IN A TENNIS GAME USING HIGH-LEVEL INFORMATION

Qiang Huang Stephen Cox

Audio, Speech and Language Processing Group
University of East Anglia

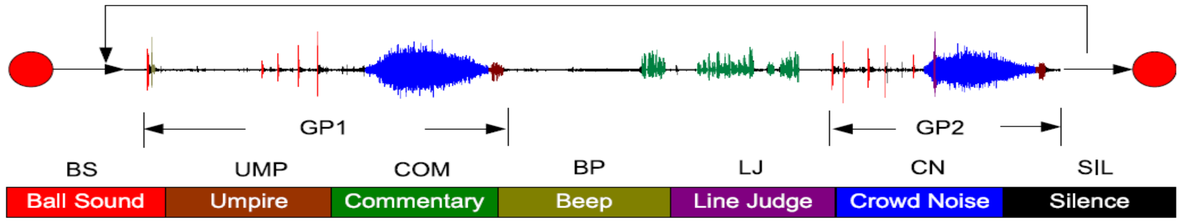
MOTIVATION

- Our ultimate goal is to combine audio and video information from a recording of a sports game (tennis) to understand the progress of the game and infer its rules, and hence make progress to a general framework for understanding events
- Here, we focus on *the audio only*. Our partners (Surrey University) are working on video processing.
- The chair umpire's speech is one of the most useful information streams, and we focus on identifying these segments of signal for later speech recognition.

OUR PREVIOUS WORK

- Seven audio events are annotated in the training set:

Event Name	Acronym	Description
Silence	SIL	Either silence, or a sound not in one of the next six categories
Umpire	UMP	Speech from the chair umpire usually current score or match progress
Commentary	COM	Commentators' speech
Line Judge	LJ	A line judge's call for a fault or a ball that is "out"
Ball Sound	BS	The sound generally by the ball striking a racquet, the ground or the net
Crowd Noise	CN	The crowd's applause, gasp, roar, etc.
Beep	BP	The electronic sound generated when the ball is out or touches the net during a serve



- **Our previous work on the detection of audio events:**
  - **Baseline:** Gaussian mixture models (GMMs) of each event
  - Incorporation of pitch and event duration information
  - Use of a hierarchical two-level "language" model [1]
  - Use of inter-event timing information [2]

PROBLEMS

- The identification of the occurrence of the chair umpire's speech is very difficult if the acoustic characteristics of chair umpires speeches differ between training and testing data (e.g. different voices, transmission channels, environmental noise etc).
- The chair umpire's speech is often corrupted by crowd noise and the commentators' voices in the game.
- The duration of chair umpire's speech is very short, generally less than 1s.

STEP 1: COARSE LOCATION OF UMPIRE SPEECH (UMP) SEGMENTS

- **Crowd noise (CN) is relatively stationary across different soundtracks, so identify all putative CN segments**
- Then use *high-level knowledge* derived from training-set to identify coarse location of UMP
  - audio event bigram model tells us that next event after CN is *most likely* to be UMP
  - duration models tell us how long UMP is likely to be
  - inter-event timing models tell us how long after CN UMP will follow
- Having found coarse position of UMP, use iterative method to refine estimate

STEP 2: REFINEMENT OF LOCATION OF UMP SEGMENTS

COM-GMM = GMM representing all COM audio events
UMP-GMM = GMM representing all UMP audio events

```
while not_converged
      foreach frame in putative UMP segment
         compute p1 = Pr(frame|COM-GMM) and p2 = Pr(frame|UMP-GMM)
            if p1/p2 < THRESHOLD
               add frame to store of frames for UMP segments
            else
               add frame to store of frames for COMM segments
            end
       end

       Re-estimate position of UMP/COMM boundary using removed frame information
       Re-estimate GMMS for UMP and COMM from frame stores

end
```

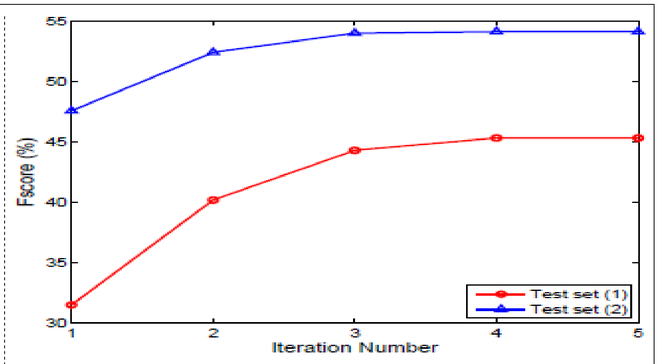
EXPERIMENTAL SET-UP

- An audio track is represented as a sequence of 39D vectors (12 MFCC coeffs. + energy +Δ + ΔΔ) estimated every 30ms.
- $\text{Fscore} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$, where $\text{Precision} = \frac{\#\text{Correctly detected umpire segments}}{\#\text{True umpire segments}}$ and $\text{Recall} = \frac{\#\text{Correctly detected umpire segments}}{\#\text{Detected umpire segments}}$

DATA & RESULTS

	Training set	Test set (1)	Test set (2)
Type	Men's single	Men's single	Women's double
#Tracks	10	3	4
Time	200 mins	150 mins	120 mins
#Comme.	3	3	1
Language Commen./ Umpire	English/ English	English/ English	Polish/ English
Gender of Commen.	male	male	male
Gender of Umpire	male	male	female

The audio tracks are extracted from 3 Wimbledon tennis games



F-score (%)	Training Set	Test set (1)	Test set (2)
Supervised 1	59.82	57.53	57.84
Supervised 2	59.82	16.55	11.49
Proposed approach	-	45.31	54.10

FUTURE WORK

- Improve the ability to handle the interference caused by crowd noise.
- Improve the ability to distinguish commentators' speech from the umpire's speech using the techniques of speech segregation.
- Use visual information (supplied by University of Surrey) to guide the audio event detection process

REFERENCES

[1] Huang, Q., Cox, S., *Hierarchical Language Modeling for Audio Events Detection in a Sports Game.* Proc. ICASSP 2010

[2] Huang, Q., Cox, S., *Using High-level Information to Detect Key Audio Events in a Tennis Game.* Proc. Interspeech 2010