Caption-aided Speech Detection In Videos

Cong Li, Zhijian Ou

ozj@tsinghua.edu.cn

Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

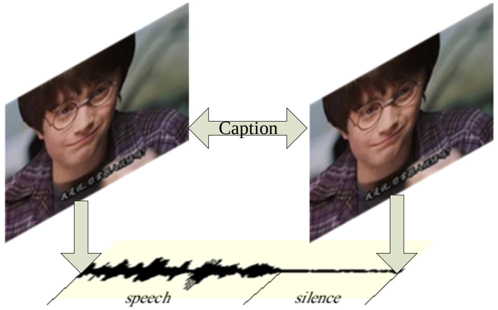
Objective: Speech detection in real world media(movies, TV series)

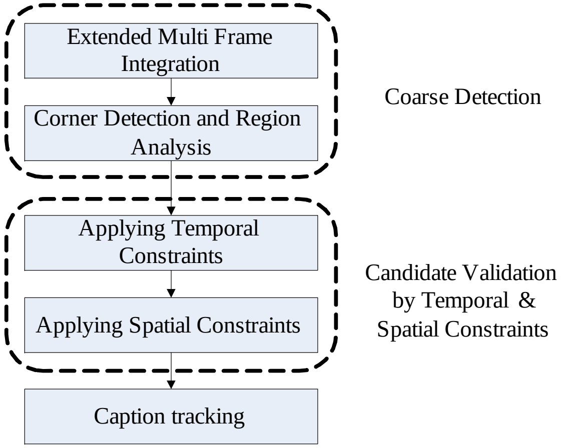
Method: Detect captions from video and pitch segments from audio
Refine results via collaboration between caption and audio information

Advantage: Information from audio and video are both made use of

Caption detection

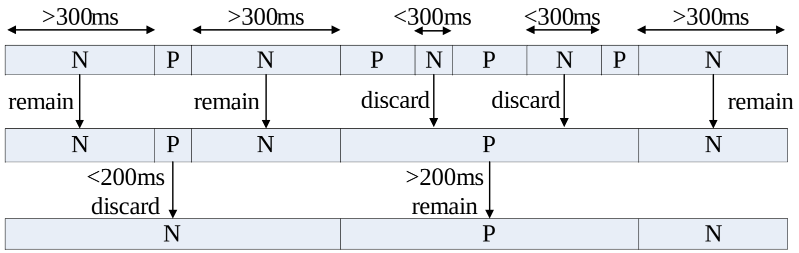
- Caption contains useful information
- But not accurate enough
- Caption detection algorithm for hard captions




Extended Multi Frame Integration
Corner Detection and Region Analysis
Coarse Detection
Applying Temporal Constraints
Applying Spatial Constraints
Candidate Validation by Temporal & Spatial Constraints
Caption tracking

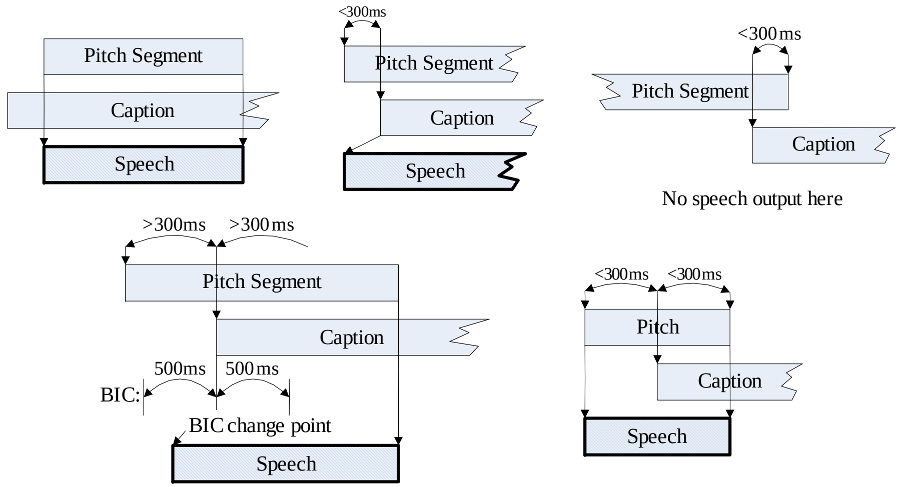
Pitch-based audio segmentation

- Pitch segments discard noise-like and silence audio
- But may include music and voice-like non speech audio
- Collaboration between caption and audio information is introduced



Smoothing of the pitch segments (plotted as 'P') and non-pitch segments (plotted as 'N')

Refinement via caption-audio collaboration



- Main idea: adjust the caption begin/end points according to the boundaries given by the pitch segments and the BIC-based change detection.
  - Pitch segments are used to eliminate the caption inaccuracy: silence, noise, laughter in caption, and temporal inaccuracy.
  - Captions are used to discard false alarm pitch segments and suggest the location of potential audio changing points.
  - BIC is introduced to determine the changing point near the potential location provided by captions.
- There are in total five specific situations to handle, as shown in the figures above.

Experiment Result

- Evaluated on 7 excerpts from 3 different TV series and 1 excerpt from a movie:

DCJ: Korean TV series "Dae-Jang-Geum", Chinese, 80 minutes.

Friends: American TV series "Friends", English, 90 minutes.

FH: Korean TV series "Full House", Korean, 30 minutes.

HP: American movie "Harry Porter and the Sorcerer's Stone", English, 30 minutes.

- Algorithm is evaluated by correct rate(correct time / time in total). A forgiveness collar of 0.25 second(both + and -) will not be scored as error.

	Friends	DCJ	FH	HP
Caption only	65.0	85.0	80.0	78.9
VAD_MLER_PR	78.4	87.8	85.0	82.8
PitchSeg_MLER	90.5	81.5	81.4	74.2
PitchSeg_BIC_MLER	90.7	89.4	86.9	85.7
Caption-aided	**92.4**	**93.0**	**96.5**	**86.6**