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

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

<table>
<thead>
<tr>
<th>Situation</th>
<th>Pitch Segment</th>
<th>Speech</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;200ms</td>
<td>N</td>
<td>P</td>
</tr>
<tr>
<td>&gt;200ms</td>
<td>P</td>
<td>N</td>
</tr>
<tr>
<td>&gt;300ms</td>
<td>P</td>
<td>N</td>
</tr>
<tr>
<td>Caption</td>
<td>N</td>
<td>P</td>
</tr>
</tbody>
</table>

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:
  - HP: American movie “Harry Potter 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.

<table>
<thead>
<tr>
<th></th>
<th>Friends</th>
<th>DCJ</th>
<th>FH</th>
<th>HP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caption only</td>
<td>65.0</td>
<td>85.0</td>
<td>80.0</td>
<td>78.9</td>
</tr>
<tr>
<td>VAD_MLER_PR</td>
<td>78.4</td>
<td>87.8</td>
<td>85.0</td>
<td>82.8</td>
</tr>
<tr>
<td>PitchSeg_MLER</td>
<td>90.5</td>
<td>81.5</td>
<td>81.4</td>
<td>74.2</td>
</tr>
<tr>
<td>PitchSeg_BIC_MLER</td>
<td>90.7</td>
<td>89.4</td>
<td>86.9</td>
<td>85.7</td>
</tr>
<tr>
<td>Caption-aided</td>
<td>92.4</td>
<td>93.0</td>
<td>96.5</td>
<td>86.6</td>
</tr>
</tbody>
</table>