Improving Boundary Estimation in Audiovisual Speech Activity Detection Using Bayesian Information Criterion

Fei Tao  
John H.L. Hansen  
Carlos Busso

Multimodal Signal Processing (MSP) Laboratory,  
Center for Robust Speech Systems (CRSS),  
Department of Electrical Engineering,  
The University of Texas at Dallas,  
Richardson TX 75080, USA
Introduction

• Speech Activity Detection (SAD) plays an important role in speech-based interfaces
• Audio-only SAD (A-SAD) may fail
  • Noise
  • Different speech mode (e.g. whisper speech)
• Introduce Visual SAD (V-SAD) to improve SAD
  [Aubrey et al. (2007), Joosten et al. (2013)]
One key problem exists in V-SAD system was the precise detection of boundaries.

- Lip movement associated with non-speech event (e.g. lip smacking, laughing)
- Anticipatory facial movements (e.g. 10 ms)
- Low video resolution (30 fps vs. 100 fps)

- Bayesian Information Criterion (BIC) to improve boundary detection
Previous Work on SAD

• Supervised V-SAD
  • Aubrey et al (2007) applied HMM in developing V-SAD system;
  • Joosten et al (2013) applied SVM classifier

• AV-SAD Fusion
  • Takeuchi et al. (2009) combined the V-SAD and A-SAD decision boundaries using logical operators.
  • Almajai and Milner (2008) concatenated acoustic and visual features.

• No one has worked on improving the boundary detection
AV-SAD System: Audio Component

• Framework proposed by Sajadi and Hansen (2013)
• Audio feature (5-D)
• Principal Component Analysis (PCA) on audio feature: 1-D combo feature

harmonicity

clarity

prediction gain

periodicity

perceptual spectral flux

Combo Feature

PCA
Unsupervised A-SAD

- Unsupervised clustering with EM approach
AV-SAD System: Video Component

- Video feature [Tao et al (2015)]:
  - Optical flow: OF\textsubscript{x}, OF\textsubscript{y} and OF\textsubscript{x}+OF\textsubscript{y} (OF\textsubscript{xy})
  - Geometric feature: height (H), width (W), W \times H and H+W
  - Short term statistics (0.3 s window)

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>OF\textsubscript{x}</th>
<th>OF\textsubscript{y}</th>
<th>OF\textsubscript{xy}</th>
<th>H</th>
<th>W</th>
<th>W+H</th>
<th>WxH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temporal Variance</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Zero Crossing Rate</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Speech Periodic Characteristic</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>First Order Derivative</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

25-D feature in total
Unsupervised V-SAD

- Similar approach to unsupervised A-SAD
- PCA on 25-D feature
- EM to form two classes on “combo” feature

25-D → PCA → Visual combo feature

![Histogram diagram showing speech and non-speech classes with a threshold between them.](image)
Proposed Approach

- Audio-visual fusion
  - Logical fusion: “AND” and “OR”
- BIC refine

Audio (5D) → PCA → EM Algorithm → Fusion → Logical “AND” → BIC Algorithm

Video (25D) → PCA → EM Algorithm → Fusion → Logical “AND” → BIC Algorithm
Bayesian Information Criterion (BIC) Refine

- The BIC is a criterion used to select a model among potential candidate models [Zhou and Hansen (2005)]
  - Hypothesis 1 (H1): one single distribution
  - Hypothesis 2 (H2): bimodal distribution
  - \( \Delta \text{BIC} = \text{BIC}(H2) - \text{BIC}(H1) \)

\[
\text{BIC}(H2) = \frac{1}{2}d \left[ N \log(2\pi) + \frac{N}{2} \log(|\hat{\Sigma}|) + \frac{N}{2} \left( N \log \left( \sum_{i=1}^{N} |\Sigma_i| \right) \right) \right] + \frac{1}{2d} \left( d \log(N) + (d + d) \log(1) \right) \cdot \log(N)
\]

\( d \) is the feature dimension
\( \hat{\Sigma} \) is covariance of N frames,
Bayesian Information Criterion (BIC) Refine

- Focus on transition area
  - Potential boundary given by previous steps
  - $\Delta$BIC computed for each frame in search window
  - Extra frames before and after search window

Speech  Non-Speech

Search Window

0.5s  0.5s

potential boundary
Bayesian Information Criterion (BIC) Refine

- Focus on transition area
  - Potential boundary given by previous steps
  - $\Delta$BIC computed for each frame in search window
  - Extra frames before and after search window

$\Delta$BIC?
Corpus Description

- MSP Audio-visual Whisper (MSP-AVW) corpus
  - 20 males and 20 females
  - 120 TIMIT sentences per speaker (60 in neutral, 60 in whisper)
  - Audio: SHURE 48 KHz close-talk microphone
  - Video: high definition SONY cameras (1440 × 1080) at 29.97 fps
Experiment and Result

- **Performance without BIC**
  - Whisper decreases performance by ~20%
  - V-SAD is robust to different modes
  - Under neutral condition, the fusion decreases the performance by ~5%
    - The ground truth of the labels was annotated based only on audio
    - Original sampling frequency is low (29.97 fps)
  - Under whisper condition, the fusion improves the performance by ~8%

<table>
<thead>
<tr>
<th>Modality</th>
<th>Set</th>
<th>Acc [%]</th>
<th>Pre [%]</th>
<th>Rec [%]</th>
<th>F [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-SAD</td>
<td>Nsen</td>
<td>94.05</td>
<td>97.15</td>
<td>89.85</td>
<td>93.35</td>
</tr>
<tr>
<td></td>
<td>Wsen</td>
<td>67.96</td>
<td>61.02</td>
<td>88.65</td>
<td>72.28</td>
</tr>
<tr>
<td>V-SAD</td>
<td>Nsen</td>
<td>78.06</td>
<td>75.11</td>
<td>89.45</td>
<td>80.40</td>
</tr>
<tr>
<td></td>
<td>Wsen</td>
<td>78.20</td>
<td>72.69</td>
<td>89.10</td>
<td>80.06</td>
</tr>
<tr>
<td>AV-SAD</td>
<td>Nsen</td>
<td>89.47</td>
<td>97.90</td>
<td>79.93</td>
<td>88.00</td>
</tr>
<tr>
<td></td>
<td>Wsen</td>
<td>81.28</td>
<td>81.73</td>
<td>79.21</td>
<td>80.45</td>
</tr>
</tbody>
</table>
- **Performance with BIC:**
  - Apply BIC on detected boundary from AV-SAD

<table>
<thead>
<tr>
<th></th>
<th>Set</th>
<th>ACC [%]</th>
<th>Pre [%]</th>
<th>Rec [%]</th>
<th>F [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AV-SAD</strong></td>
<td>Nsen</td>
<td>89.47</td>
<td>97.90</td>
<td>79.93</td>
<td>88.00</td>
</tr>
<tr>
<td></td>
<td>Wsen</td>
<td>81.28</td>
<td>81.73</td>
<td>79.21</td>
<td>80.45</td>
</tr>
<tr>
<td><strong>AV-SAD</strong></td>
<td>Nsen</td>
<td>91.11</td>
<td>97.47</td>
<td>83.77</td>
<td>90.10</td>
</tr>
<tr>
<td>+ A-BIC</td>
<td>Wsen</td>
<td>82.91</td>
<td>84.47</td>
<td>79.48</td>
<td>81.90</td>
</tr>
<tr>
<td><strong>AV-SAD</strong></td>
<td>Nsen</td>
<td>88.53</td>
<td>92.22</td>
<td>83.18</td>
<td>87.47</td>
</tr>
<tr>
<td>+ V-BIC</td>
<td>Wsen</td>
<td>78.67</td>
<td>76.63</td>
<td>80.54</td>
<td>78.53</td>
</tr>
<tr>
<td><strong>AV-SAD</strong></td>
<td>Nsen</td>
<td>91.25</td>
<td>97.49</td>
<td>84.05</td>
<td>90.27</td>
</tr>
<tr>
<td>+ AV-BIC</td>
<td>Wsen</td>
<td>82.87</td>
<td>83.76</td>
<td>80.37</td>
<td>82.03</td>
</tr>
</tbody>
</table>

- **A-BIC improves the system:**
  - For speech detection, ~2% absolute improvement
- **V-BIC impairs the system**
  - Modalities mismatch
- **AV-BIC achieves best performance on speech detection**
Median Local Boundary Mismatch

- Local Boundary Mismatch (LBM)
  - the mismatch frames between the detected boundary and ground truth in local regions

- Median Local Boundary Mismatch (MLBM)
  - Represents the boundary detection performance
  - Lower is better
Boundary detection performance:

- Up-sampling to 100 fps for MLBM comparison

<table>
<thead>
<tr>
<th>Set</th>
<th>MLBM [fps]</th>
</tr>
</thead>
<tbody>
<tr>
<td>AV-SAD</td>
<td></td>
</tr>
<tr>
<td>Nsen</td>
<td>35.00</td>
</tr>
<tr>
<td>Wsen</td>
<td>64.00</td>
</tr>
<tr>
<td>AV-SAD + A-BIC</td>
<td></td>
</tr>
<tr>
<td>Nsen</td>
<td>25.00</td>
</tr>
<tr>
<td>Wsen</td>
<td>56.00</td>
</tr>
<tr>
<td>AV-SAD + V-BIC</td>
<td></td>
</tr>
<tr>
<td>Nsen</td>
<td>42.00</td>
</tr>
<tr>
<td>Wsen</td>
<td>71.00</td>
</tr>
<tr>
<td>AV-SAD + AV-BIC</td>
<td></td>
</tr>
<tr>
<td>Nsen</td>
<td>25.00</td>
</tr>
<tr>
<td>Wsen</td>
<td>53.00</td>
</tr>
</tbody>
</table>

- A-BIC improves the system:
  - For MLBM, relatively improve 28.5% under neutral and 12.5% under whisper

- V-BIC impairs the system
  - Modalities mismatch

- AV-BIC achieves best performance on boundary detection
Conclusion and Future Work

• Conclusion
  • AV-SAD is explored showing that visual modality will improve robustness under whisper condition
  • Proposed a approach to improve boundary detection in SAD by BIC
  • AV-BIC achieves best performance

• Future Work
  • Better fusion approach need be explored
THANK YOU!

QUESTION?