An Unsupervised Visual-only Voice Activity Detection Approach Using Temporal Orofacial Features

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
Outline

• Introduction
• Related Work
• Corpus Description
• Proposed Approach
• Experiment and Result
• Conclusions and Future Work
Introduction

• Voice Activity Detection (VAD) plays an important role in speech-based interfaces

• Audio based VAD (AVAD) has challenges:
  • Background noise
  • Different speech modes (e.g. emotion, soft speech, whisper)

• Visual VAD (VVAD) becomes an alternative
Introduction

• Voice Activity Detection (VAD) plays an important role in speech-based interfaces

• Audio based VAD (AVAD) has challenges:
  • Background noise
  • Different speech modes (e.g. emotion, soft speech, whisper)

• Visual VAD (VVAD) becomes an alternative
Related Work

**Supervised:**

- Navarathna et al. [2011] extracted discrete cosine transform coefficients around mouth and augment them by their derivative.
- Aubery et al. [2007] used active appearance model and retinal filter to detect speech activity based on HMM
- Takeuchi et al. [2009] extracted the variance of optical flow as visual features and proposed audiovisual VAD system
Related Work (Cont.)

Unsupervised:

• Sodoyer et al. [2006] proposed an unsupervised method to detect lip activity by adopting a threshold.

• Sadjadi and Hansen [2013] proposed a state-of-the-art unsupervised approach for AVAD

Benefit:

• No training data

• Adapt to testing conditions

• Unsupervised approach offers more flexibility
Corpus Description

- Audio-visual Whisper (AVW) corpus
- 20 males and 20 females
- Corpus consists of
  - Digits
  - Read sentence (120 TIMIT sentences: 60 in neutral and 60 in whisper)
  - Spontaneous talk
- Audio collected with a SHURE 48 KHz close-talk microphone
Corpus Description

- Video collected with high definition SONY cameras (1440 × 1080) at 29.97 fps (label based on audio)
Proposed Approach

- Video processing and facial feature extraction
- Estimation of dynamic and temporal features
- Principle component analysis (PCA)
- Expectation maximum (EM) algorithm for clustering

**Original Feature**
- Optical Flow Variance x
- Optical Flow Variance y
- Height
- Width

**EM Cluster**

**Analysis Flow**
- Original
- ST-ZCR
- ST-VAR
- Delta
Feature Extraction

• 66 landmarks detected by CSIRO [Cox et al., 2013]
• Quality check with the outputs from another system
• Orofacial feature extraction:
  • height(H) and width(W)
  • variance of optical flow in x direction (OFx) and y direction (OFy)
Dynamic and Temporal Features

• Facial feature vector (7D):
  • Overall optical flow variance (OF_{xy}): OF_x + OF_y
  • Overall distance (H + W) & approximate area (H × W)

• Statistics over facial feature vector
  • Dynamic features
    • Delta: first order difference
  • Temporal features over 7D vector:
    • ST-VAR: short-term (0.3s) variance
    • ST-ZCR: short-term zero-crossing rate

Original Feature
Optical Flow Variance x
Optical Flow Variance y
Height
Width

Original
ST-ZCR
ST-VAR
Delta
PCA
EM Cluster
Dynamic and Temporal Features

- **Facial feature vector (7D):**
  - Overall optical flow variance \((\text{OF}_{xy})\): \(\text{OF}_x + \text{OF}_y\)
  - Overall distance \((H + W)\) & approximate area \((H \times W)\)

- **Statistics over facial feature vector**
  - Dynamic features
    - Delta: first order difference
  - Temporal features over 7D vector:
    - ST-VAR: short-term (0.3s) variance
    - ST-ZCR: short-term zero-crossing rate

**Original Feature**
- Optical Flow Variance \(x\)
- Optical Flow Variance \(y\)
- Height
- Width

**EM Cluster**

**Diagrams and Flowchart**

- Original
- ST-ZCR
- ST-VAR
- Delta
- PCA
- EM Cluster
Dynamic and Temporal Features

- **Facial feature vector (7D):**
  - Overall optical flow variance ($OF_{xy}$): $OF_x + OF_y$
  - Overall distance ($H + W$) & approximate area ($H \times W$)

- **Statistics over facial feature vector**
  - Dynamic features
    - Delta: first order difference
  - Temporal features over 7D vector:
    - ST-VAR: short-term (0.3s) variance
    - ST-ZCR: short-term zero-crossing rate

<table>
<thead>
<tr>
<th>Original Feature</th>
<th>Original</th>
<th>ST-ZCR</th>
<th>ST-VAR</th>
<th>Delta</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optical Flow Variance x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Optical Flow Variance y</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Height</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Width</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**EM Cluster**
Feature Set

- Final feature vector consists of 19 features

<table>
<thead>
<tr>
<th>Set</th>
<th>$OF_x$</th>
<th>$OF_y$</th>
<th>$OF_{xy}$</th>
<th>$H$</th>
<th>$W$</th>
<th>$H+W$</th>
<th>$H \times W$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Delta*</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>ST-VAR*</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>ST-ZCR*</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

- ST-ZCR: short term zero crossing rate;
- ST-VAR: short term variance;
- Delta: first order difference
Unsupervised Classification

- Principle component analysis (PCA) applied on final feature to form a 1-D combo feature
- Inspired by Sadjadi and Hansen [2013]
Unsupervised Classification

- Principle component analysis (PCA) applied on final feature to form a 1-D combo feature
- Expectation maximum (EM) algorithm is run for clustering
Baseline AVAD

- Audio only VAD (proposed by Sadjadi and Hansen [2013]):
  - 5D feature: Harmonicity, Clarity, Prediction Gain, Periodicity, Perceptual Spectral Flux
  - Changing speech mode impair the system performance (20% drop)

<table>
<thead>
<tr>
<th>Set</th>
<th>Precision[%]</th>
<th>Recall[%]</th>
<th>F-score[%]</th>
<th>Accuracy[%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutral</td>
<td>91.3</td>
<td>98.0</td>
<td>94.5</td>
<td>93.9</td>
</tr>
<tr>
<td>Whisper</td>
<td>78.7</td>
<td>72.3</td>
<td>75.3</td>
<td>74.8</td>
</tr>
</tbody>
</table>

\[
F - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}
\]
Experiment and Results

- Video only VAD (proposed approach):
  - Visual cues are robust to different speech modes
  - For neutral sentence, the performance is about 13% lower than AVAD system
  - For whispered sentence, the performance is about 6% higher than AVAD system

<table>
<thead>
<tr>
<th>Set</th>
<th>Precision[%]</th>
<th>Recall[%]</th>
<th>F-score[%]</th>
<th>Accuracy[%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutral</td>
<td>90.7</td>
<td>73.8</td>
<td>81.4</td>
<td>80.0</td>
</tr>
<tr>
<td>Whisper</td>
<td>90.3</td>
<td>73.5</td>
<td>81.1</td>
<td>79.4</td>
</tr>
</tbody>
</table>
Compare AVAD and VVAD

- Anticipatory movement of lips
- Lower resolution for visual modality
Compare Supervised and Unsupervised

- Training set: 20 speakers; testing set: 20 speakers
- Unsupervised setting:
  - Proposed approach is applied on the testing data
- Supervised setting:
  - Linear kernel SVM built with training set

| Training: 20 Spkrs | Testing: 20 Spkrs |
Compare Supervised and Unsupervised

- Training set: 20 speakers; testing set: 20 speakers
- Unsupervised setting:
  - Proposed approach is applied on the testing data
- Supervised setting:
  - Linear kernel SVM built with training set

Testing: 20 Spkrs
Compare Supervised and Unsupervised

- Training set: 20 speakers; testing set: 20 speakers
- Unsupervised setting:
  - Proposed approach is applied on the testing data
- Supervised setting:
  - Linear kernel SVM built with training set
Compare Supervised and Unsupervised

- Training set: 20 speakers; testing set: 20 speakers
- Unsupervised setting:
  - Proposed approach is applied on the testing data
- Supervised setting:
  - Linear kernel SVM built with training set

<table>
<thead>
<tr>
<th>Set</th>
<th>Supervised VVAD</th>
<th>Unsupervised VVAD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P[%]</td>
<td>R[%]</td>
</tr>
<tr>
<td>Neutral</td>
<td>89.1</td>
<td>84.3</td>
</tr>
<tr>
<td>Whisper</td>
<td>88.7</td>
<td>84.2</td>
</tr>
</tbody>
</table>
Benefits of Supervised Approach

• Supervised approach is 5% higher than unsupervised approach
  • Trade-off

• Unsupervised approach is 5% higher when tested on a different corpus

• Benefits of supervised approach is gone
Conclusions and Future Work

• A new unsupervised VVAD approach is proposed
• The proposed approach is robust to speech mode changing
• Audiovisual VAD will be explored in future to improve the performance under the neutral mode
Acknowledgments

- Thanks to National Science Foundation (NSF)

Reference: