**Background**

Speech provides an attractive modality for human-machine interface (HMI) through automatic speech recognition (ASR). But ASR suffers from three primary limitations:

1. Performance degradation in presence of ambient noise
2. Limited ability for privacy/secrecy
3. Poor accessibility for those with speech disorders.

These limitations have motivated the need for alternative non-acoustic modalities of subvocal or silent speech recognition (SSR).

**Objective**

We set out to design and develop a SSR system based on recordings of the surface electromyographic (sEMG) signal from articulator muscles of the face and neck during silently mouthed (subvocal) speech.

**Methods**

1) **Experiment Setup**

- **Subjects** — n=18 healthy (11 males, 8 females, age range 18-42 y.o.)
- **sEMG Sensors**: 8 DE 2.1 sensors and Trigno™ Mini sensors (Delsys, Natick, USA)
- **Protocol** — Subjects silently mouthed words while sEMG activity was recorded from muscles of face and neck.

2) **Data Collection**

**Data Corpus**

<table>
<thead>
<tr>
<th>Data Corpus</th>
<th>Subjects</th>
<th>Vocabulary/Phrases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Isolated Words</td>
<td>Controls (n=9)</td>
<td>65 words and digits</td>
</tr>
<tr>
<td>Sequences of Words</td>
<td>Controls (n=4)</td>
<td>202 words, 1,200 sequences</td>
</tr>
<tr>
<td>Continuous Speech</td>
<td>Controls (n=5)</td>
<td>2,200 words, 1,200 phrases</td>
</tr>
</tbody>
</table>

3) **Initial Data Processing**

- Separating speech from non-speech sEMG activity
- Finite multi-channel state machine
- Robust against noise

**Algorithm Development – Strategic evolution of Hidden Markov Models (HMMs) for SSR**

**Challenge 1**

Discriminating isolated words using sEMG-based speech related features

**Challenge 2**

Tracking sequences of words from patterns of sEMG signals using grammatical context

**Challenge 3**

Recognizing a large vocabulary of untrained words using phoneme-based models

**Final sEMG SSR System – Algorithms, Sensors and Mobile Deployment**

**SSR System Configuration:**

- **Sensors** — sEMG 4 sensor-array (under development) worn on face and neck
- **Features** — MFCCs
- **Grammar** — NL Equivalence
- **Recognition Toolkit** — KALDI
- **Model** — HMM Triphone, HLDA Feature Reduction, maximum likelihood linear regressions (MLLR), subspace Gaussian mixture modelling (SGMM)

**SSR Sensors: Trigno™ Quadruo Facial Array**

- Wireless/Bluetooth communication for mobile use

**SSR System Performance (WER):**

<table>
<thead>
<tr>
<th>Subject</th>
<th>Digits</th>
<th>Text</th>
<th>Special Operations</th>
<th>Common Phrases</th>
<th>Mean WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.7</td>
<td>0.9</td>
<td>2.0</td>
<td>0.0</td>
<td>1.4</td>
</tr>
<tr>
<td>2</td>
<td>15.4</td>
<td>15.9</td>
<td>8.0</td>
<td>15.4</td>
<td>13.9</td>
</tr>
<tr>
<td>3</td>
<td>20.7</td>
<td>12.1</td>
<td>13.6</td>
<td>5.2</td>
<td>12.9</td>
</tr>
<tr>
<td>4</td>
<td>18.2</td>
<td>10.3</td>
<td>13.6</td>
<td>5.2</td>
<td>10.7</td>
</tr>
<tr>
<td>5</td>
<td>12.2</td>
<td>5.6</td>
<td>3.4</td>
<td>1.2</td>
<td>5.6</td>
</tr>
</tbody>
</table>

**Mean WER**

- **Mean WER**: 8.9%

**Conclusion**

- Our SSR system was able to recognize subvocal speech with 8.9% WER from a 2,200-word vocabulary of 1,200 continuous phrases including previously unseen words.
- The miniaturized sensors provide a robust and unencumbering facial interface that can transmit data via custom wireless or Bluetooth protocol for portable integration with a mobile device.
- These results demonstrate the viability of our SSR system as a silent modality of speech communication that can be developed further for persons with speech impairments (Meltzner et al, 2017), military personnel, or consumer applications.

**Acknowledgements**

- VocaliD, Inc. Belmont, USA
- MGH, Boston, USA
- BAE Systems, Inc. Burlington, USA

**References**