SYSTEMS FOR ROBUST SPEECH ACTIVITY DETECTION AND THEIR RESULTS WITH THE RT05 AND RT06 EVALUATION TESTS

Dušan Macho, Andrey Temko, Climent Nadeu
TALP Research Center, UPC, Barcelona

IV Jornadas de Tecnologías del Habla, Zaragoza, November 8-10th, 2006
Speech Activity Detector (SAD) for interactive meeting/lecture smart-room scenario

CHIL project:
- unobtrusive technologies and demos: far-field microphone setup, single and multiple mics
- online technologies: low-delay and real-time
- Aimed to several technologies (SID, SLOC, AED, ASR) and services

What has been added to our baseline SAD system:
- Three new SAD features in addition to LDA-FF features
- Two alternative classifiers have been tested in addition to Decision Tree: GMM and SVM

Submitted to NIST RT06 evaluation. Training of the SVM version was posteriorly enhanced to cope with the problems of that classifier in our application
Features

**Idam**

30ms/10ms frame length/shift

**Frequency Filtering (FF) of FBE:** filter \( h(k) = \{1, 0, -1\} \Rightarrow \text{static FF} \)

**Time derivatives:** \( \Delta FF, \Delta \Delta FF, \Delta \log E \) appended to static FF \((16+16+16+1=49)\)

**LDA:** 49-element FF vector is reduced to 1-element scalar Idam

**Ifed, hfed**

\[
E_i(t) = \log \left( \sum_k S(k, t) \right)
\]

where \( k \) correspond to 0.4-1.2kHz and 4.5-6.5 kHz for Ifed and hfed, respectively

\[
dE_1(t) = \frac{1}{60} \sum_{i=-4}^{4} i \cdot E_1(t+i) \quad \text{Ifed}(t) = \frac{1}{5} \sum_{i=-2}^{2} \text{abs}(dE_1(t+i))
\]
Feature Extractor

\[
\begin{align*}
&\text{FF} \\
&\Delta\text{FF} \\
&\Delta\Delta\text{FF} \\
&\Delta\log E \\
&\text{lfed} \\
&\text{hfed}
\end{align*}
\]

\[
\text{Feature vector} = \text{ldam}(t-15, t-10, t-6, t-3, t, t+3, t+6, t+10), \text{lfed}(t), \text{hfed}(t), \text{xfed}(t)
\]

\[
\text{xfed}(t) = \frac{1}{2} \times (\sqrt{[\text{hfed}(t-9) \times \text{lfed}(t+9)]} + \sqrt{[\text{hfed}(t+9) \times \text{lfed}(t-9)]})
\]
Features, cont.

- Signal
- Spec-gram
- Spec-gram
- Speeched
- Spectral
- Spectrum
- Spectrum
- Spectrogram
## Classifier

### Gaussian Mixture Model (GMM)
- **32 mixtures** for both Speech and Non-Speech with **diagonal covariance matrix**
- **20 iterations of EM algorithm** for Gaussian mixture model training
- **Classifier used in systems submitted for both** “confmtg” and “lectmtg” tasks

### Support Vector Machine (SVM)
- **Training data set reduced** to 20 thousand feature vectors by random selection
- **Gaussian kernel; parameters set via 5-fold cross-validation** on the reduced training data
RT06 Post-processing

sdm
- 11 frame majority voting along time
- Addition of 0.2s at the beginning and the end of each speech segment

mdm
- sdm SAD for each channel (without post-proc.)
- Majority voting for each frame using info from several channels
- 11 frame majority voting along time
- Addition of 0.2s at the beginning and the end of each speech segment
Training data

Training used for LDA and the classifier

- confmtg: SPEECON and RT05 meeting
- lectmtg: above and CHIL

<table>
<thead>
<tr>
<th>Database</th>
<th>SPEECON</th>
<th>RT05 meetings</th>
<th>CHIL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language</td>
<td>Spanish</td>
<td>English</td>
<td>English</td>
</tr>
<tr>
<td>Type</td>
<td>Single utterances</td>
<td>Meeting</td>
<td>Lecture</td>
</tr>
<tr>
<td>Microphone</td>
<td>2-3m in front of speaker</td>
<td>On the table</td>
<td>On the table</td>
</tr>
<tr>
<td>Signal</td>
<td>16 kHz, 16 bit</td>
<td>16 kHz, 16 bit</td>
<td>16 kHz, 16 bit</td>
</tr>
</tbody>
</table>
Performance

**Metrics**
*(RT05 and RT06)*

**NIST** = \( \frac{\text{Duration of Incorrect Decisions}}{\text{Duration of All Speech}} \)

**Other metrics**
*(CHIL)*

**MR** = \( \frac{\text{Duration of Incorrect Decisions}}{\text{Duration of All (Speech and Non-Speech)}} \)

**SDER** = \( \frac{\text{Missed Speech}}{\text{Duration of All Speech}} \)

**NDER** = \( \frac{\text{Missed Non-Speech}}{\text{Duration of All Non-Speech}} \)
## Performance, cont.

**RT 05**

- **A**: ldam (8 features)
- **B**: ldam + lfed (9 features)
- **C**: ldam + hfed (9 features)
- **D**: ldam + hfed (10 features)

<table>
<thead>
<tr>
<th>NIST</th>
<th>MR / SDER / NDER</th>
<th>Feat set</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>20.69 / 18.77 / 24.32</td>
<td>12.37 / 11.20 / 30.51</td>
<td>14.76 / 13.37 / 30.27</td>
<td>11.54 / 10.43 / 33.42</td>
</tr>
<tr>
<td><strong>SVM</strong></td>
<td></td>
<td></td>
<td>12.25 / 11.13 / 33.81</td>
<td><strong>8.47</strong> / <strong>7.69</strong> / <strong>38.42</strong></td>
<td>10.02 / 9.11 / 47.70</td>
<td>8.66 / 7.88 / 49.00</td>
</tr>
<tr>
<td><strong>GMM</strong></td>
<td></td>
<td></td>
<td><strong>...</strong></td>
<td><strong>...</strong></td>
<td><strong>...</strong></td>
<td><strong>...</strong></td>
</tr>
</tbody>
</table>
### Performance, cont.

#### RT 06

<table>
<thead>
<tr>
<th>GMM</th>
<th>NIST MR / SDER / NDER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Idam + Ifed + hfed + xfed (11 features)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>confmtg</td>
</tr>
<tr>
<td>sdm</td>
<td>5.45</td>
</tr>
<tr>
<td></td>
<td>5.1 / 3.1 / 41.4</td>
</tr>
<tr>
<td>mdm</td>
<td>5.63</td>
</tr>
<tr>
<td></td>
<td>5.3 / 3.5 / 38.7</td>
</tr>
</tbody>
</table>
• Two modifications
  – Efficient sample selection using the fast training PSVM, and selecting 20 chunks of 1000 samples that show the highest 5-fold cross-validation accuracies
  – Penalize more the Speech class (as NIST metric does) by introducing different costs for the two classes
| Method                  | NIST  
|------------------------|--------
|                        | MR / SDER / NDER |
| GMM                    | 8.47   |
|                        | 7.69 / 4.61 / 38.42 |
| SVM                    | 11.45  |
|                        | 10.41 / 7.99 / 34.56 |
| SVM enhanced training  | 8.03   |
|                        | 7.30 / 2.51 / 55.07 |
Conclusions

- **SAD** oriented towards smart room environments and evaluated in NIST RT05 and RT06
- **Significant improvement** by using additional features and either GMM or SVM classifiers
  - Error reduction from 20.69% to 8.47% with GMM in RT05
  - Further reduction (8.03%) with better trained SVM
- **Competitive results in RT06**