Speech Detection for Text-Dependent Speaker Verification

Orith Toledo-Ronen
Persay Ltd.
Outline

• Motivation
• Review of existing techniques
• HMM-based speech detection
• The Evaluation Track corpus
• Experimental results
• Summary
Motivation

• Improving end-point detection improves text-dependent speaker verification performance
• Existing algorithms: energy-based voice activity detector (VAD)
• Problem: background speech may pass the energy threshold
Existing Techniques

• Energy
• Amplitude
• Zero-crossing rate
• Linear prediction error
• Pitch
• HMM
Comparison of Techniques

• Energy-based VAD
  - Statistics on frame energy
  - Threshold setting

• HMM-based VAD
  - Speaker dependent model
  - Password detection
  - Filters the noise
Energy-based VAD

- Compute the energy of all frames
- Find statistics of energy values $\Omega(E)$
- Compute the energy threshold
  \[ T = f(\Omega(E)) \]
- Filter out all frames with energy below $T$
HMM-based VAD

- A left-to-right hidden Markov model of the phrase
- *Not* phoneme-based
- Trained from 3 repetitions
Training

• Use the energy-based VAD first
• Train the speaker HMM
• Train a background HMM from:
  - noise segments
  - background speech
• Merge the speaker and background HMMs
Merging Models

Audio

Noise Speaker Noise

Noise
Speaker
Noise
Detection

• Run Viterbi with the merged HMM and find the speaker’s states in the segmentation

• Use the HMM VAD as a filter before verification
Example

Audio

Energy VAD

HMM VAD

Energy+HMM VAD
The Evaluation Track Corpus

• **Database**: Persay’s TD corpus

• **Passwords**: 9-digit telephone number
  4-digit personal code

• **Speakers**: 45 males
  37 females

• **Impostors**: up to 5 same-gender impostors for each speaker
The Evaluation Track Corpus

• **Sessions:** ~5 calls per speaker with 3 repetition of each password in each call
• **Media:** cellular phone
• **Language:** Hebrew
Experimental Results

- **Results**: % Equal Error Rate

<table>
<thead>
<tr>
<th>Gender</th>
<th>Password</th>
<th>Energy</th>
<th>HMM</th>
<th>H+E</th>
<th>E+H</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>9-digit</td>
<td>7.2</td>
<td>8.1</td>
<td>8.7</td>
<td>6.7</td>
</tr>
<tr>
<td></td>
<td>4-digit</td>
<td>11.1</td>
<td>12.6</td>
<td>10.8</td>
<td>9.0</td>
</tr>
<tr>
<td>Female</td>
<td>9-digit</td>
<td>6.3</td>
<td>5.8</td>
<td>7.1</td>
<td>6.4</td>
</tr>
<tr>
<td></td>
<td>4-digit</td>
<td>10.8</td>
<td>12.2</td>
<td>12.5</td>
<td>12.4</td>
</tr>
</tbody>
</table>
Password Rejection

• **Impostor**: the Viterbi path does not reach the speaker’s model
• **Partial password**: the Viterbi path does not cover all the speaker’s states

<table>
<thead>
<tr>
<th>Gender</th>
<th>Password</th>
<th>H+E</th>
<th>E+H</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>9-digit</td>
<td>1 / 39</td>
<td>5 / 54</td>
</tr>
<tr>
<td></td>
<td>4-digit</td>
<td>0 / 21</td>
<td>3 / 45</td>
</tr>
<tr>
<td>Female</td>
<td>9-digit</td>
<td>2 / 52</td>
<td>6 / 82</td>
</tr>
<tr>
<td></td>
<td>4-digit</td>
<td>1 / 33</td>
<td>7 / 68</td>
</tr>
</tbody>
</table>

% Rejected
(Target / Impostor)
Password Rejection - Cont’d

• The Persay’s TD corpus was manually cleaned by a human listener.
• Rejected by human: 102 target attempts 115 impostor attempts
• Algorithm rejection: 33% target attempts 86% impostor attempts
Password Rejection - Cont’d

• Segments rejected by human and algorithm:
  - non-speech: DTMFs, ring tone, silence
  - corrupted audio
  - wrong password
  - strong background speech

• Segments rejected only by human:
  - all contain the password, by poor quality
  - low volume, background speech, error and repair
Summary

• We have presented a method for speech detection in a text-dependent speaker verification system.
• The HMM-based VAD can be used in combination of an energy-based VAD.
• It can detect the password and reject invalid verification audio segments.