AN EXPLORATION ON INFLUENCE FACTORS OF VAD’S PERFORMANCE IN SPEAKER RECOGNITION

Cheng Gong, CSLT
2013/04/15
Outline

• Introduction
• Analysis about influence factors of VAD’s performance
• Experimental results and analysis
• Conclusions
Introduction

- Voice activity detection (VAD)
  - A method for detecting periods of speech in observed signals
  - VAD technique is particularly important and widely used in both automatic speech recognition and speaker recognition
- Two parts of VAD process:
  - Acoustic feature extraction
  - Decision mechanism
- Currently used VAD methods:
  - Short-term signal energy, zero-crossing rate
  - Speech/noise spectral characteristics based methods:
    - MFCCs, LTSE, LSF, MMSE, etc.
  - Periodic feature based methods:
    - ACF, F0, etc.
  - ......
Introduction

- Difficulties of VAD:
  - Determine end-points accurately
  - Be robust to noise, especially to non-stationary noise

- Basic principle of choosing end-points:
  - Speech recognition: integrity of the speech contents
  - Speaker recognition: typicality of the speaker characteristics

To get a better result, VAD method in speaker recognition may be different from which in speech recognition
Phonation Types

- Voiced sound
  - Glottis excitation + Vocal Tract response
  - Quasi-periodic signal
  - All simple/compound vowels and 4 initial consonants (m, n, l, r) in mandarin are voiced sound

- Unvoiced sound
  - No vocal cord vibration
  - Non-periodic signal
  - Plosive/affricate/fricative, aspirated/unaspirated
  - The other initial consonants in mandarin are unvoiced sound
Phonation Types’ Influence

- Research Assumption
  - Phonation types’ distinction may lead to different contribution to speaker verification results

- Research Procedure
  - To segment the speech signals based on phonation types (Using HVITE tools)
  - To splice the speech according to the rules of classification by person
    - Silence segments
    - Voiced sound segments
    - Unvoiced sound segments
  - To extract features (MFCC), train models and test on the speaker verification system, compare and analyse the results.
SNR’s Influence

• Research Assumption
  • Noise in the speech doesn’t reflect the speaker’s characteristics, so the parts which has low SNR may lead to a high EER

• Research Procedure
  • To estimate noise power spectrum of each speech signal
  • To Calculate SNR of each frame
  • To splice the speech based on the SNR level
    • ‘Clean’~20dB, 20dB~15dB, 15dB~10dB, 10dB~ 5dB, 5dB~
  • To extract features (MFCC), train models and test on the speaker verification system, compare and analyse the results.
Noise Estimation Algorithm

- Analysis object
  - Additive noise in the speech (stable/unstable)

- Destination
  - To obtain a noise power spectrum estimation from noisy speech

- Implement method
  - Combination of minimum statistics, continuous spectral minimum & minima controlled recursive algorithm
Compute smooth speech power spectrum $P(\lambda, k)$

Find the local minimum of noisy speech $P_{\text{min}}(\lambda, k)$

Compute ratio $S_r(\lambda, k)$ of smoothed speech power spectrum to its local minimum

Calculate speech presence probability $p(\lambda, k)$ using first-order recursion

Compute time-frequency dependent smoothing factors $\alpha_s(\lambda, k)$

Update noise estimate $D(\lambda, k)$ using time-frequency dependent smoothing factors $\alpha_s(\lambda, k)$
Database

- CCB database
  - Recorded in clean environments using telephone channel
  - Sampling rate: 8kHz
  - Training utterance length: 39s~75s
  - Testing utterance length: 11s~44s

<table>
<thead>
<tr>
<th>Channel</th>
<th>Training</th>
<th>True Speaker</th>
<th>Impostor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>F</td>
<td>M</td>
</tr>
<tr>
<td>Telephone</td>
<td>50</td>
<td>50</td>
<td>150</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td></td>
<td>300</td>
</tr>
</tbody>
</table>
Experimental Conditions

- **Feature:**
  - Mel-frequency cepstral coefficients (MFCC)
  - 16-orders with energy, without delta
  - 32 Mel filter banks

- **Model:**
  - GMM-UBM
  - 1024 mixtures
## Results and Analysis

### Phonation types’ influence

<table>
<thead>
<tr>
<th>Gender</th>
<th>Voiced Sound</th>
<th>Unvoiced Sound</th>
<th>Silence</th>
<th>Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>7.65</td>
<td>42.49</td>
<td>48.74</td>
<td>8.17</td>
</tr>
<tr>
<td>F</td>
<td>8.13</td>
<td>42.17</td>
<td>49.22</td>
<td>8.53</td>
</tr>
<tr>
<td>M+F</td>
<td>5.89</td>
<td>41.87</td>
<td>49.12</td>
<td>7.44</td>
</tr>
</tbody>
</table>
Results and Analysis

If the segments (SNR < 5dB) are removed, the EER = 5.09% (the baseline EER = 7.16%)
Results and Analysis

- Add white noise with different SNRs to the speech, the table below shows the EERs when removing the segments whose SNR < 5dB:

<table>
<thead>
<tr>
<th>SNR</th>
<th>EER(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Improved</td>
</tr>
<tr>
<td>clean</td>
<td>5.09</td>
</tr>
<tr>
<td>20dB</td>
<td>6.46</td>
</tr>
<tr>
<td>15dB</td>
<td>8.31</td>
</tr>
<tr>
<td>10dB</td>
<td>11.85</td>
</tr>
<tr>
<td>5dB</td>
<td>16.65</td>
</tr>
</tbody>
</table>
Conclusion

- Unvoiced sounds don’t contribute much to speaker verification results. The speech with voiced sounds only can get better results.
- The EER is related to SNR of the speech directly, if we remove some segments whose SNR is very low, the results will get much better. This method has remarkable effects on noisy speech.
References


References


Thank you!