FORMANT-GAPS FEATURES FOR SPEAKER VERIFICATION USING WHISPERED SPEECH

Abinay Reddy Nani, Achuth Rao M V and Prasanta Kumar Ghosh
Electrical Engineering, Indian Institute of Science (IISc), Bangalore, India-560 012

Introduction

- **Speaker verification(SV):** To verify whether a given test speech recording is from an enrolled speaker or not.
- **Whisper:** Used in private conversations, pathological conditions.
- **Need for whisper SV:** Speakers often whisper the password in a biometric system, criminals might whisper in phone to avoid leaving the voice print.[1]
- **Challenges:** Absence of pitch, Low-frequency formant shift, hyper-articulation

Whispered speaker verification system

Proposed Formant-Gaps features

- For each frame, we computed five formants using [3], indicated by a vector of 
  \[ F = [f_1, f_2, f_3, f_4, f_5] \]  
  where \( f_i \) indicates the \( i \)-th formant. Let us consider first \( f_1 \) and second order \( f_1^2 \) formant gaps \((F_{CG})\) as
  \[ f_1^2 = f_1 - f_0, \quad 1 \leq i \leq 4, \quad f_1^2 = f_1 - f_1^2, \quad 1 \leq i \leq 3 \]  
  Let \( F^1 = [f_1^1; 1 \leq i \leq 4], F^2 = [f_1^2; 1 \leq i \leq 3] \).
- We experimented two features using \( F_{CG} \)s, namely, 
  \( F_{CG} = [F^1, F^2] \) and \( F_{CG} = [F^1, F^2, F^3] \).
- Illustrative experiment:

In order to understand the distribution of the proposed features, we trained a speaker specific GMM for whispered and neutral speech features separately.

**Data(N|W): The KL divergence between i-th speaker’s neutral GMM (\( N_i \)) and whispered GMM (\( W_i \)).**

**MKL(i): The average of KL divergence between \( N_i \) and \( W_i \) speakers.**

\[
M_{KL}(i) = \frac{1}{N_i} \sum_{j=1}^{N_i} D(N_i|W_j) \quad \sigma_{KL}(i) = \frac{1}{\sqrt{N_i}} \sum_{j=1}^{N_i} (D(N_i|W_j) - M_{KL}(i))^2
\]

where \( P^* = \{i : D(N_i|W_i) < M_{KL}(i) - 1.5 \times \sigma_{KL}(i)\} \).

Experiments & Results

- **Data set:** We considered data from 3 databases (CHAINS,wTIMIT,TIMIT) with 714 speakers comprising 29232 neutral and 22932 whispered recordings.
- **Baseline features:**
  - **MFCC:** 13-dimensional mel frequency cepstral coefficients along with velocity and acceleration coefficients to make 39 dimensional features.
  - **AAMF:** Auditory-inspired amplitude modulation features (40-dimensional)[4].
  - **DNN:** Deep neural network(DNN) based feature mapping on both MFCC and AAMF features are considered.

- **Equal error rate(ER)** for different test conditions:

<table>
<thead>
<tr>
<th>Table: EER using proposed and baseline features</th>
<th>Table: EER with varying number of whisper recordings (( N_w )) in enrollment</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>f</strong></td>
<td><strong>Test condition</strong></td>
</tr>
<tr>
<td>-------</td>
<td>--------------------</td>
</tr>
<tr>
<td><strong>F</strong></td>
<td>5</td>
</tr>
<tr>
<td><strong>F</strong></td>
<td>9</td>
</tr>
<tr>
<td><strong>F</strong></td>
<td>12</td>
</tr>
<tr>
<td><strong>F</strong></td>
<td>4</td>
</tr>
<tr>
<td><strong>F</strong></td>
<td>8</td>
</tr>
</tbody>
</table>

- **Features mapping on the baseline feature:** (MFCC and AAMF), when only neutral data used in enrollment and testing using whispered speech.

- **The SV using baseline features requires at least four whisper recordings in the enrollment phase for it to perform better than the proposed features.**

Conclusion

- **We experimented formant-gaps based features for whispered speaker verification.**

The experiments revealed that the proposed features are robust to the modes (whisper and neutral) of speech for SV applications.

- **Future work :** Experimenting with different feature mapping methods for whispered speaker verification.

References