FORMANT-GAPs FEATURES FOR SPEAKER VERIFICATION USING WHISPERED SPEECH

Abinay Reddy Naini, 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 speech: 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 ($f_1^1$) and second order ($f_1^2$) formant gaps ($F_{1G}$) as
|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 ($f_1^1$) and second order ($f_1^2$) formant gaps ($F_{1G}$) as
$f_1^1 = f_{i+1} - f_{i-1}, 1 \leq i \leq 4, 
$ $f_1^2 = f_{i+1} - f_{i-1}, 1 \leq i \leq 3$ (1)\n|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 $F_{1G}$, namely,
|$F_{1G} = [F^1, F^2]$ and $F_{2G} = [F^1, F^2, F^3].$
|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.
|$D(N|W)$: The KL divergence between the $i$-th speaker’s neutral GMM ($N_i$) and whispered GMM ($W_i$).
|where the dimension of features $F_{1G}, F_{2G}$ are 9,12 respectively.

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 (EER) for different test conditions:

<table>
<thead>
<tr>
<th>Features</th>
<th>Test condition</th>
<th>whisper</th>
<th>Neutral</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F^1$</td>
<td>22.42</td>
<td>6.28</td>
<td></td>
</tr>
<tr>
<td>$F_{1G}$ (9)</td>
<td>13.00</td>
<td>7.8</td>
<td></td>
</tr>
<tr>
<td>$F_{1G}$ (12)</td>
<td>14.98</td>
<td>9.14</td>
<td></td>
</tr>
<tr>
<td>MFCC (39)</td>
<td>22.47</td>
<td>6.25</td>
<td></td>
</tr>
<tr>
<td>AAMF (40)</td>
<td>19.81</td>
<td>4.4</td>
<td></td>
</tr>
<tr>
<td>MFCC (40)</td>
<td>17.01</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>AAMF (40)</td>
<td>16.79</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

Table: EER with varying number of whisper recordings ($N_e$) in enrollment

<table>
<thead>
<tr>
<th>$N_e$</th>
<th>AAMF (40)</th>
<th>FoG</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>17.01</td>
<td>13.00</td>
</tr>
<tr>
<td>2</td>
<td>14.14</td>
<td>10.82</td>
</tr>
<tr>
<td>4</td>
<td>8.61</td>
<td>9.68</td>
</tr>
<tr>
<td>6</td>
<td>6.14</td>
<td>8.88</td>
</tr>
<tr>
<td>8</td>
<td>4.78</td>
<td>8.46</td>
</tr>
</tbody>
</table>

The combination of $F$ and $F_i$ features ($F_{1G}$) performs the best, when only neutral data used in enrollment and tested using whispered speech.

The feature mapping on the baseline feature (MFCC DNN and AAMF DNN) performs better compared to (MFCC and AAMF), when only neutral data used in enrollment and tested 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 proposed 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.

Acknowledgement: Authors thank the Pratiksha Trust for their support.

References


Table: EER with varying number of whisper recordings ($N_e$) in enrollment