Line Spectral Frequencies

- LSF = alternative representation of LPC
- position of particular line spectral frequency (LSF) => shape of spectral envelope
Main idea

LSFs shifting ~ spectral envelope transformation

=> shifting by frequency axis warping
LSF warping function (1)

• clustering of joint LSF vector $z = [x^T, y^T]^T$ into $K$ classes (bisective k-means)
  - mean vector
    $$\mu_z^k = \begin{bmatrix} \mu^k_x[1], \ldots, \mu^k_x[P], \mu^k_y[1], \ldots, \mu^k_y[P] \end{bmatrix}^T$$
  - covariance matrix
    $$\Sigma_z^k = \text{diag} \begin{bmatrix} \sigma^k_x[1], \ldots, \sigma^k_x[P], \sigma^k_y[1], \ldots, \sigma^k_y[P] \end{bmatrix}$$
  - warping function
    $$\tilde{y}[i] = f_k(x[i])$$

• resulting conversion function
  $$\tilde{y}[i] = \sum_{k=1}^K w_k(x) f_k(x[i])$$
LSF warping function (2)

- requirements for warping function of k-th class
  \[ f_k(\mu^k_x[j]) = \mu^k_y[j] \]
- in each interval
  \[ \langle \mu^k_x[j], \mu^k_x[j+1] \rangle \]
  warping function is partly linear (WL function)
  \[ \tilde{y}[i] = a^j_k x[i] + b^j_k \]
  or partly cubic (WC function)
  \[ \tilde{y}[i] = a^j_k x^3[i] + b^j_k x^2[i] + c^j_k x[i] + d^j_k \]
LSF warping function (3)

Comparison of warping functions

- WC function
- WL function
- $[\mu_x, \mu_y]$
Experiments

- voice conversion system
  - true envelope estimator => LSFs
    - GMM-based transformation
    - warping function
  - pitch transformed by Gaussian normalisation
  - spectral detail transformed by residual prediction
- training set – 40 short sentences
- testing set – 10 sentences
Objective evaluation (1)

- performance index
  \[ P_{LSF} = 1 - \frac{\sum_{n=1}^{N} d(\tilde{y}_n, y_n)}{\sum_{n=1}^{N} d(x_n, y_n)} \]
- higher value => better similarity

<table>
<thead>
<tr>
<th>Function</th>
<th>Number of classes</th>
<th>Target speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>M1</td>
</tr>
<tr>
<td>GMM</td>
<td>5</td>
<td>0.424</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.425</td>
</tr>
<tr>
<td>WC</td>
<td>5</td>
<td>0.235</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.264</td>
</tr>
<tr>
<td></td>
<td>150</td>
<td>0.314</td>
</tr>
</tbody>
</table>
Objective evaluation (2)

- global variance ratio \( R_{GV} = \frac{1}{P} \sum_{p=1}^{P} \frac{GV(\tilde{y}[p])}{GV(y[p])} \)

- value closer to 1 => better quality

<table>
<thead>
<tr>
<th>Function</th>
<th>Number of classes</th>
<th>Target speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>M1</td>
</tr>
<tr>
<td>GMM</td>
<td>5</td>
<td>0.642</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.655</td>
</tr>
<tr>
<td>WC</td>
<td>5</td>
<td>1.081</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>1.038</td>
</tr>
<tr>
<td></td>
<td>150</td>
<td>0.935</td>
</tr>
</tbody>
</table>
Subjective evaluation

- listening test – 10 participants
- 20 quintuples of utterances:
  - source, target, 3 transformed
- average order (lower value => better similarity / quality)
- GMM => better similarity
- WC => better quality

<table>
<thead>
<tr>
<th></th>
<th>GMM</th>
<th>combined</th>
<th>WC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Similarity</td>
<td>1.83 ± 0.56</td>
<td>1.94 ± 0.31</td>
<td>2.23 ± 0.59</td>
</tr>
<tr>
<td>Quality</td>
<td>2.34 ± 0.55</td>
<td>2.02 ± 0.45</td>
<td>1.65 ± 0.62</td>
</tr>
</tbody>
</table>
Thank you for your attention.