**Goal**
- Many voice conversion methods produce muffled synthesized outputs due to over-smoothing of the converted spectra.
- GV enhancement – used for muffling reduction and commonly applied as an integrated part of the conversion system.
- We propose a new modular method for GV enhancement, applied as a post-processing block.

**Voice Conversion**
- Transform a sentence said by a source speaker, to sound as if a target speaker had said it, based on pre-recorded training set.

**Proposed GV Enhancement**
- GV enhancement methods have been proposed to overcome the muffling effect:
  - ML estimation [Toda et al., 2007]
  - Constrained GMM (CGMM) [Benisty and Malah, 2011]
- These enhancement methods are integrated into the training process of the conversion.

**GV Enhancement Using GMM**
- Linear Conversion based on a Gaussian Mixture Model (GMM) [Stylianou, 1998], [Kain & Macon, 1998]
- A common approach for spectral conversion.
- Minimizes the mean Log Spectral distortion (LSD) between converted feature vectors and target vectors.
- Characterized by smoothed spectral envelopes causing a muffled effect.

**GV Enhancement Using an LSD Constraint**
- Input
  - A sequence of converted feature vectors $\tilde{Y}_{tr} = (y_1, y_2, \ldots, y_T)^\top$
- Output
  - A sequence of enhanced feature vectors $Z_{tr} = (z_1, z_2, \ldots, z_T)^\top$
- The enhanced sequence is the solution of:
  $Z_{tr} = \arg \max_{z_{tr}} \text{NGV}(Z_{tr})$
  subject to $\text{LSD}(Z_{tr}, \tilde{Y}_{tr}) \leq \theta_{LSD}$
- $\text{NGV}(Z_{tr})$ - the normalized GV of the sequence $Z_{tr}$, evaluated by:
  $\text{NGV}(Z_{tr}) = \frac{1}{P} \frac{\text{Var}[\{Z_{tr}(p)\}]}{\text{Var}[\{Y_{tr}(p)\}]}$
- $\text{LSD}(Z_{tr}, \tilde{Y}_{tr})$ - mean Log Spectral Distortion between the converted and enhanced sequences.
- $\theta_{LSD}$ - pre-set threshold value for the mean LSD in dB.

**Experimental Results**
- **Objectively**
  - For a given mean LSD, CGMM leads to higher GV than our method.
- **Subjectively**
  - Our method was selected by the majority of listeners as better than CGMM, both in terms of quality and similarity to the target speaker.

---

Subjective Evaluations

Quality - AB

Identity - XAB

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean LSD (dB)</th>
<th>Mean Norm. GV</th>
</tr>
</thead>
<tbody>
<tr>
<td>LS-GMM</td>
<td>6.2</td>
<td>0.1</td>
</tr>
<tr>
<td>Enhanced $\theta = 1 \text{dB}$</td>
<td>6.4</td>
<td>0.2</td>
</tr>
<tr>
<td>Enhanced $\theta = 2 \text{dB}$</td>
<td>6.7</td>
<td>0.3</td>
</tr>
<tr>
<td>Enhanced $\theta = 4 \text{dB}$</td>
<td>7.3</td>
<td>0.4</td>
</tr>
<tr>
<td>CGMM</td>
<td>7.3</td>
<td>0.9</td>
</tr>
</tbody>
</table>

Enhanced vs. Conversion, Identity, and Quality Score [Stylianou, 1998]