**Transforming F0 Contours**

**Ben Gillett, now with Camel Audio (www.camelaudio.com)**

**Simon King, Centre for Speech Technology Research, University of Edinburgh, UK**

ben@camelaudio.com, Simon.King@ed.ac.uk

---

**Introduction**

To transform the F0 contour of some speech from a source speaker, such that listeners believe it to have been uttered by some target speaker.

---

**Goal**

To transform the F0 contour of some speech from a source speaker, such that listeners believe it to have been uttered by some target speaker.

---

**Applications**

- Voice transformation
- Also need a method for transforming voice quality - see poster by King & Gillett in session PWeBe
- Speech synthesis
- As a way of adapting existing intonation models (trained on one speaker) to a new speaker, without having to annotate much data.

---

**Mapping**

**Standard method**

Usual method of normalising F0 is to use this mapping:

\[ M_N(x) = \frac{(x - \mu_N)}{\sigma_N} \times \sigma_{\text{target}} + \mu_{\text{target}} \]

**New method**

A nonlinear mapping, \( M_{PL} \), composed of piecewise linear sections between F, L, H and S.

---

**Evaluation**

**Perceptual experiment**

25 subjects presented with speech from target speaker and speech with transformed F0 contours (“imitator speech”) in XABX format. Asked to judge which of A or B was most like X. A and B varied between: actual target F0, normalised F0 (standard method), transformed F0 (new method).

Speaker pairs classified as “similar” (\( S_{\text{same}} \)) or “different” (\( S_{\text{different}} \)).

**Results**

<table>
<thead>
<tr>
<th></th>
<th>Mean (%)</th>
<th>Std. Dev.</th>
<th>Value</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preference for ( M_{PL} ) over ( M_N ) for ( S_{\text{same}} )</td>
<td>67</td>
<td>10</td>
<td>(&lt; 1 \times 10^{-4})</td>
<td>-8.71</td>
</tr>
<tr>
<td>Preference for ( M_{PL} ) over ( M_N ) for ( S_{\text{different}} )</td>
<td>54</td>
<td>8</td>
<td>(~0.01)</td>
<td>-2.49</td>
</tr>
<tr>
<td>Preference for target over mapped contours</td>
<td>73</td>
<td>9</td>
<td>(&lt; 1 \times 10^{-4})</td>
<td>-13.8</td>
</tr>
</tbody>
</table>

---

**After Patterson**

- Sentence-initial high (S)
- Non-initial accent peaks (H)
- Post-accent valleys (L)
- Sentence-final low (F)

---

**What next?**

Apply the method to full voice transformation or speech synthesis.

---

**See also...**

- Poster by Gillett & King in session PWeBe (voice quality transformation)
- www.cstr.ed.ac.uk for latest progress on voice transformation and speech synthesis
- www.camelaudio.com for musical instrument transformation and morphing