Automatic transformation of irregular to regular voice by residual analysis and synthesis

Tamás Gábor Csapó, Géza Németh, (csapot.nemeth)@tmit.bme.hu

1. Introduction
• Creaky voice, laryngealization, vocal fry, glottalization
• Irregular vibration of vocal folds
• Irregular F0 and/or amplitudes (see Fig. 1)
• Occurrence: up to 15% of vowels of natural speech
• Phrase boundaries
• Sentence endings
• Vowel-vowel transitions
• Differences compared to regular speech (see Fig. 1)
• Time between successive glottal pulses longer and more irregular
• Lower F0 and higher jitter

2. Transformation method
• Effect of irregular voice on pitch estimation
  • Causes errors in standard methods (see Fig. 2, b)
  • Interpolation necessary (see Fig. 2, c)
• Continuous pitch tracking (CONT_F0) [2]
  • Standard autocorrelation
  • No voiced/unvoiced decision
  • Kalman smoothing-based interpolation
• Irregular-to-regular transformation (see Fig. 4)
  • Based on our analysis-synthesis vocoder [3]
• Codebook of pitch synchronous residuals
• Parameters: F0, gain: frame by frame energy
  • R0: prominent values in the frame (see Fig. 5)
• Dynamic modeling
• MGC: Mel-Generalized Cepstrum
• Automatic irregular voice detection using [5]
• Result: quasi-regular speech signal (see Fig. 1, 3)

3. Perceptual evaluation
• Stimuli for the perceptual evaluation
  • Four Hungarian speakers (three males: FF1, FF3, FF4 and one female: NO3)
• Five sentences from each speaker, containing irregular voice in at least 15% of the utterances
• Irregular sections were transformed to modal voice by the proposed method
• Web-based listening test: “roughness” and “naturalness” of samples
• Both versions of each sentence (original and transformed to regular), altogether 40 uterances (4 speakers x 5 sentences x 2 versions)
• Two 5-point MOS-like questions:
  • Roughness (“1 – very rough”, 5 – not rough at all”)
  • Naturalness (“1 – very unnatural… 5 – very natural”)
• All listeners (11 males, 2 females), native speakers of Hungarian, university students or speech technology experts; on average 8 minutes to complete
• Results of the listening test (see Table 1)
• For speakers FF1, FF3 and FF4, the method was able to decrease the perceived roughness (only significant for speaker FF4)
• For speaker NO3, the transformation slightly increased the roughness
• Naturalness slightly decreased (significant for FF1, FF3 and NO3)

4. Discussion and Conclusions
• Full automatic method to transform irregular voice to regular voice
  • Codebook-based residual analysis-synthesis
  • Original irregular sections replaced by overlap-added frames from codebook
  • More suitable than direct waveform manipulation like PSOLA with handcrafted weights, as the residual can be corrected automatically
• Kalman smoothing of CONT_F0: more suitable than the simple linear F0 interpolation we used in [1]
• In some cases it causes high F0 at the end of the sentence (unnatural)
• Solution might be: combine F0 interpolation with rule-based intonation model
• It is known that several types of glottalization can be differentiated
• Our method was suitable for transforming the type used by speaker FF4
• Applications of the model may include
  • Correction of voices where unwanted irregular phonation occurs frequently (e.g. those of radio announcers or voice actors)
  • Transform glottalized parts of large speech databases (help further automatic speech processing; voice conversion)

Key references

Table 1. Speaker by speaker blanks and standard deviations (in parentheses) for the roughness and naturalness questionnaires.

<table>
<thead>
<tr>
<th>speaker</th>
<th>original roughness</th>
<th>original naturalness</th>
<th>transformed roughness</th>
<th>transformed naturalness</th>
</tr>
</thead>
<tbody>
<tr>
<td>FF1</td>
<td>2.77 (0.93)</td>
<td>3.71 (1.03)</td>
<td>2.91 (0.93)</td>
<td>3.86 (1.03)</td>
</tr>
<tr>
<td>FF3</td>
<td>2.80 (1.28)</td>
<td>3.26 (1.08)</td>
<td>2.99 (1.08)</td>
<td>3.71 (1.08)</td>
</tr>
<tr>
<td>FF4</td>
<td>2.89 (1.05)</td>
<td>3.42 (1.10)</td>
<td>3.12 (1.09)</td>
<td>3.71 (1.09)</td>
</tr>
<tr>
<td>NO3</td>
<td>3.37 (1.13)</td>
<td>3.69 (1.18)</td>
<td>3.37 (1.18)</td>
<td>3.86 (1.18)</td>
</tr>
</tbody>
</table>

Acknowledgements
This research is partially supported by the Swiss National Science Foundation via the joint research project (SCOPES scheme) SP2: SCOPES project on speech processing (SNF n° 157320/152485-1).