Effect of formant and F0 discontinuity on perceived vowel duration: Impacts for concatenative speech synthesis

Tomáš Bořil, Pavel Šturm, Radek Skarnitzl & Jan Volín

I. Summary

In the vowels of Czech synthetic speech, we found the effect of discontinuities in formant contours ↔ perceived duration
- Unit selection concatenative synthesis systems
- Discontinuities at the concatenation point of two diphones
- Stricter penalizations of formant discontinuities in vowel concatenation would seem beneficial

Vowel quantity is contrastive in Czech language
- toto nemá [ˈtoto nemaː] means you don’t have this
- toto nema [ˈtoto nemaː] means you don’t have this

Typical audible artifacts in synthetic speech
- Errors in the database
- Imperfect correlation of the target and join costs with human perception
- Preference of low global cost over low local cost

II. Method

Material
4 target sentences (7 syllables and 3 stress groups) in a male voice using the ARTiC synthesis system (Articial Talker in Czech)

Target context = final vowel /akoʊ/ preceded by /ː/ or /ʌ/ and followed by /ŋ/

Tenhle dopis je pro vás. /ˈ provoʊs/ [This letter is for you.]
Nejdříve rozmotej prová. /pronˈoʊv/ [First disentangle the rope.]
Byl tam veliký prová. /pronˈoʊv/ [There was heavy traffic.]
Zítra natrhej synč. /ˈsɪŋtʃ/ [Tomorrow pick some currant.]

1. Duration of target vowels: PSOLA-modified (pitch synchronous overlap-add), between typical values of short and long vowels given phrase-final lengthening (see Table 1) → resynthesized to maintain the same audio quality as the manipulated stimuli → original stimuli
2. Target manipulations (see Table 2) performed on the second half of the vowel, i.e., from concatenation point (see Figure 1)
Formant manipulations: LPC Burg method (resampled to 16 kHz, prediction order of 15, window length of 25 ms, time step of 5 ms and pre-emphasis filter starting at 50 Hz)
Additional duration or F0 shifts: PSOLA
3. For the sentence with /æ/γ/, another manipulation based on 3b (−11.5% F2) but in the entire portion of the vowel (to decide: effect due to a discontinuity in formant contours or to a general shift in vowel quality?)
4. Distractors (easy items to process) and training session stimuli also included

Note: all manipulations performed in Praat

Table 1: Duration of the final vowel of original stimuli, initial and final F0 values of the second half of the vowel. F1 and F2 at the end of the first half.

<table>
<thead>
<tr>
<th>/prová.</th>
<th>/prová.</th>
<th>/promˈaʊv/</th>
<th>/ˈrɪbiːs/</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration</td>
<td>145 ms</td>
<td>132 ms</td>
<td>152 ms</td>
</tr>
<tr>
<td>F0in</td>
<td>105 Hz</td>
<td>85 Hz</td>
<td>85 Hz</td>
</tr>
<tr>
<td>F0out</td>
<td>92 Hz</td>
<td>76 Hz</td>
<td>80 Hz</td>
</tr>
<tr>
<td>F1</td>
<td>688 Hz</td>
<td>614 Hz</td>
<td>529 Hz</td>
</tr>
<tr>
<td>F2</td>
<td>1280 Hz</td>
<td>1159 Hz</td>
<td>967 Hz</td>
</tr>
</tbody>
</table>

Table 2: Performed manipulations.

<table>
<thead>
<tr>
<th></th>
<th>1a / 1b</th>
<th>2a / 2b</th>
<th>3a / 3b</th>
<th>4a</th>
<th>4b</th>
<th>5a / 5b</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>F0 shifted by +2 ST / −2 ST</td>
<td>F1 shifted by +11.5% / −11.5%</td>
<td>F2 shifted by +11.5% / −11.5%</td>
<td>F3 shifted by +11.5% / −11.5%</td>
<td>Duration shifted by +30 ms / −30 ms</td>
</tr>
<tr>
<td>1a / 1b</td>
<td>F0 shifted by +2 ST / −2 ST</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2a / 2b</td>
<td>F1 shifted by +11.5% / −11.5%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3a / 3b</td>
<td>F2 shifted by +11.5% / −11.5%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4a</td>
<td>F3 shifted by +11.5% / −11.5%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4b</td>
<td>- 4a, but in addition F0 shifted by −2 ST</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5a / 5b</td>
<td>Duration shifted by +30 ms / −30 ms</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

III. Results & Conclusions

Despite identical duration of the compared stimuli, vowels manipulated in the second part towards centralized values (i.e., less peripheral) were systematically considered to be shorter, and vice versa (see Figure 2)

- A relaxed articulatory setting in the vocalic space may be interpreted as the offset of the vowel, and the listener would then interpret the whole vowel as shorter than it really is
- However, the influence seems to be distinct from an overall formant change (without a discontinuity), see the control stimulus in 3b

In Czech, there is a vowel quality difference between [ɪ] × [iː] and the duration ratio is much lower than in the other pairs
- Duration as a cue is therefore less important
- The perceptual integration of formants F2 and F3 in [ɪː] might also possibly affect manipulations of F2 which may not be sufficient to change the position of the effective formant

No clear effect of F0 discontinuity was found

See also
Connect Praat, Matlab, and R with mPraat and rPraat
http://fu.ff.cuni.cz/praat/

Contact: borilt@gmail.com

Institute of Phonetics, Charles University
http://fonetika.ff.cuni.cz/