

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

Tomáš Bořil, Pavel Šturm, Radek Skarnitzl & Jan Volín

CONTACT: BORILT@GMAIL.COM



## I. Summary

In the vowels of Czech synthetic speech, we found the effect of discontinuities in formant contours  $\leftrightarrow$  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 nemaž* [ˈnemaʃ] means *don't erase this*
- *toto nemáš* [ˈnemaːʃ] 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 (Artificial Talker in Czech)

Target context = final vowel /a: a o i:/, preceded by /v/ or /b/ and followed by /s/

Tenhle dopis je pro <b>v</b> ás.	/ˈprova:s/	[This letter is for you.]
Nejdřív rozmotej <b>pro</b> vaz.	/ˈprovas/	[First disentangle the rope.]
Byl tam veliký <b>pro</b> voz.	/ˈprovos/	[There was heavy traffic.]
Zítřka natrhej <b>ry</b> bíz.	/ˈribi:s/	[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)  $\rightarrow$  resynthesized to maintain the same audio quality as the manipulated stimuli  $\rightarrow$  '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 /a:/, 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.

	/ˈprova:s/	/ˈprovas/	/ˈprovos/	/ˈribi:s/
Duration	145 ms	132 ms	152 ms	159 ms
F0 <sub>initial</sub>	105 Hz	85 Hz	85 Hz	103 Hz
F0 <sub>final</sub>	92 Hz	76 Hz	80 Hz	94 Hz
F1	688 Hz	614 Hz	529 Hz	259 Hz
F2	1280 Hz	1159 Hz	967 Hz	2109 Hz

Table 2: Performed manipulations.

1a / 1b	F0 shifted by +2 ST / -2 ST
2a / 2b	F1 shifted by +11.5% / -11.5%
3a / 3b	F2 shifted by +11.5% / -11.5%
4a	F1 shifted by -11.5%, F2 by +11.5%, excepting /ˈribi:s/ F1 +11.5%, F2 -11.5%
4b	~ 4a, but in addition F0 shifted by -2 ST
5a / 5b	Duration shifted by +30 ms / -30 ms

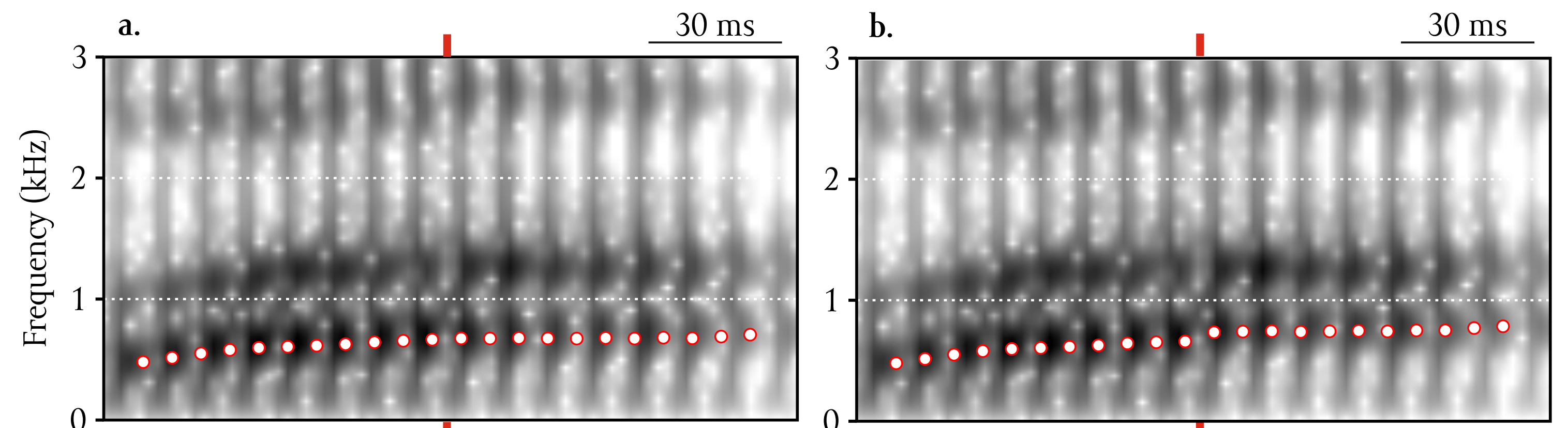


Figure 1: Comparison of an original [a:] vowel (a.) and an F1+ shift (b.) in /ˈprova:s/.

### Participants

- 25 respondents (18 females, 7 males, median age = 24)
- Native speakers of Czech, studied phonetics at Charles University
- The purpose of the experiment was not known to the participants except for 'improving the speech synthesis system'

### Test procedure

- A quiet room, headphones
- 2AFC (two-alternative forced choice) experiment in Praat
- A sequence of two phrases (one of them manipulated)
- The participants decided in which of the two phrases they thought the last syllable was longer (they could replay the item three times)
- Each target item appeared twice (in *orig* > *manip* and *manip* > *orig* order)
- All items in the test session randomized for each individual
- 5 blocks of 20 items and one block of 10 items
- 2 minutes of music between the blocks for relaxation
- Total test duration approximately 30 min.

## 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]  $\times$  [i:] and the durational 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 [i:] 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

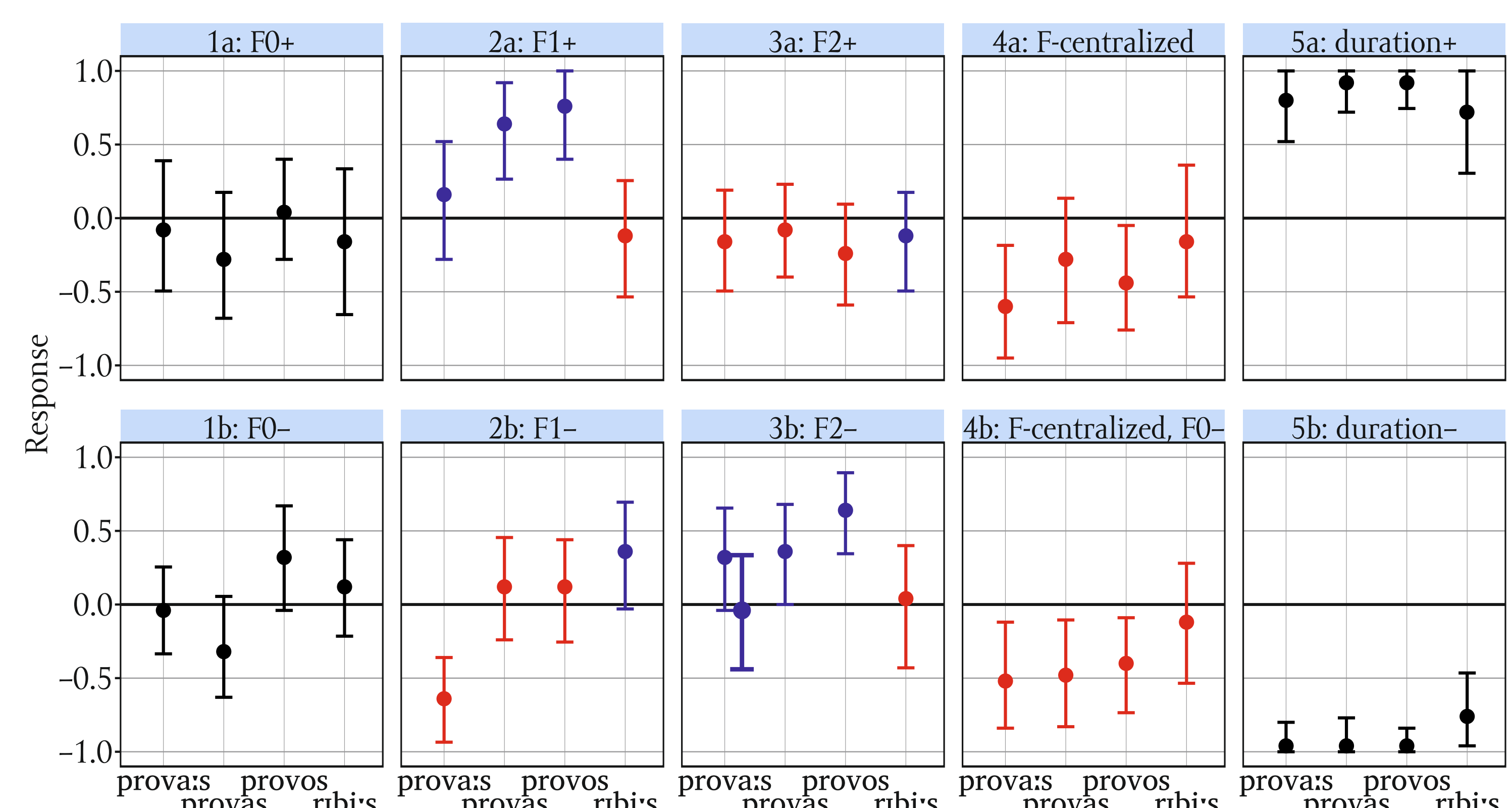


Figure 2: Mean values and conf. intervals ( $\alpha = 0.05$  with Bonferroni correction) of perceived vowel duration (+1 stands for longer, -1 for shorter perceived duration of the manipulated vowel). Blue = manipulation towards peripheral values, red = towards central values. In 3b, the bold item = manipulation on the entire portion of the vowel.