Perceived length of Czech high vowels in relation to formant frequencies evaluated by automatic speech recognition

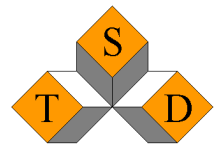
23rd International Conference on Text, Speech and Dialogue (TSD 2020)

Tomáš Bořil and Jitka Veroňková

Tomas.Boril@ff.cuni.cz

Institute of Phonetics, Charles University

FACULTY OF ARTS
Charles University



Introduction

- Czech vowel system: 5 pairs of short and long monophthongs + 3 diphthongs
- The vowel length is phonologically distinctive
- Improper interchange in L2 speakers' production may lead to misunderstanding
- [kruciː farmaːr̝i] vs [kruːciː farmaːr̝i]
  (meaning *cruel farmers* vs *turkey farmers*)
- [vɪrɪ] vs [viːrɪ] (meaning *viruses* vs *vortices*)

Introduction

Czech high vowels: perceived vowel length is influenced by formant values related to a tongue position[1].

- Bohuslav Hála (1941): [i] and [iː] differ in formant frequencies
- Jana Dankovičová (1997): IPA symbols differentiation [ɪ], [iː]
- Václav Jonáš Podlipský et al. (2009): Bohemians (the western region) – rely more on the spectrum and Moravians (the eastern region) – rely more on the duration
- Šárka Šimáčková et al. (2012): formant shifts between [u] and [uː] in production (the Bohemian region)
- Václav Jonáš Podlipský et al. (2019): formant values play crucial role in a perceptual discrimination between [u] and [uː]
  - A listening test with artificial one-syllable pseudowords

[1] The majority of naive L1 users is not aware of such relations.

Experiment 1

- [ɪ] vs [iː], real-world two-syllable words
- Minimal pair: [vɪrɪ] (meaning *viruses*) and [viːrɪ] (meaning *vortices*)
- Manipulations of the 1st syllable (21 formant steps × 7 duration steps = 147 items) in Praat and rPraat using LPC
  - 1st syllable is not prone to phrase-final lengthening
  - [ɪ] F1–F3 = 405, 2295, 2866 Hz
  - [iː] F1–F3 = 305, 2700, 3000 Hz
  - Duration 90 ms to 300 ms

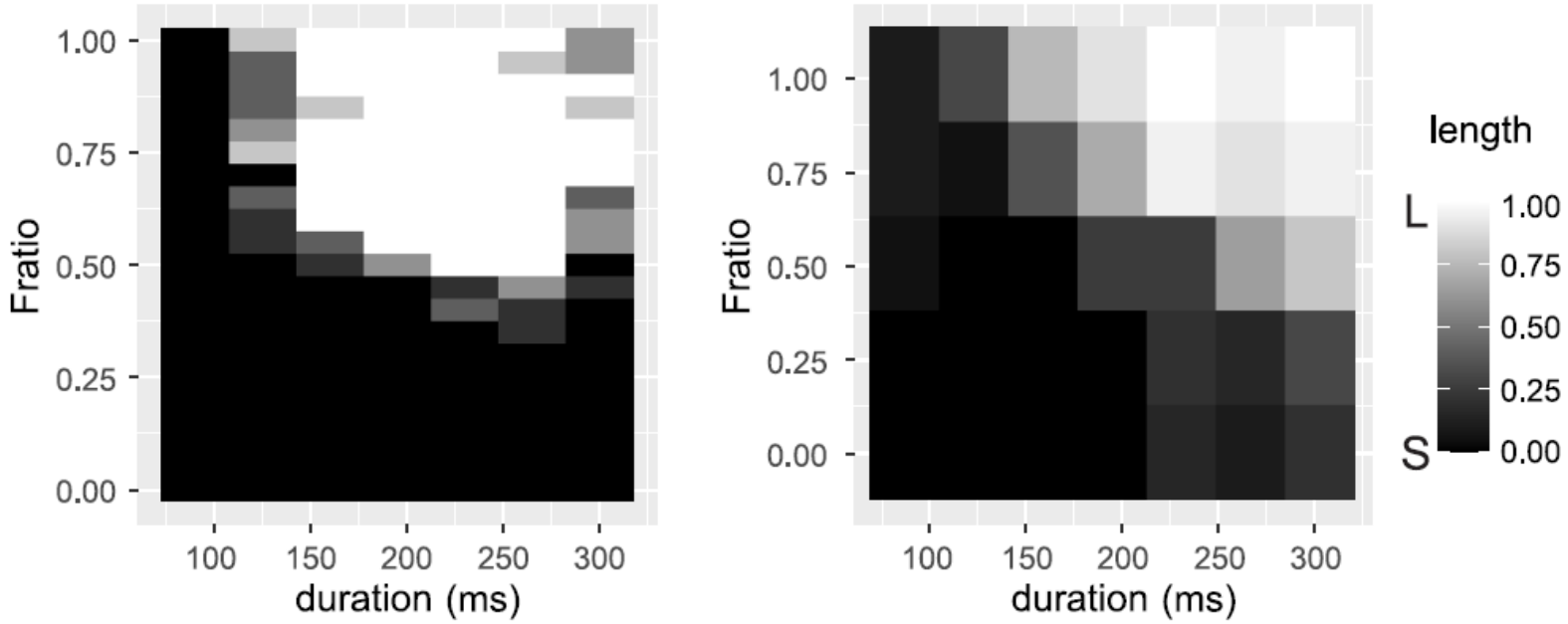
Fratio=1, dur=90ms | Fratio=1, dur=195ms | Fratio=1, dur=300ms

Fratio=0.5, dur=90ms | Fratio=0.5, dur=195ms | Fratio=0.5, dur=300ms

Fratio=0, dur=90ms | Fratio=0, dur=195ms | Fratio=0, dur=300ms

Experiment 1

- ASR: 5 replications of random permutation of items
  - Beey by NEWTON Technologies
- Listening test (human perception experiment)
  - 20 participants (Bohemian region, male and female, median age = 23 years)
  - To keep them focused, a subset of items only (5 formant steps $\times$ 7 durations = 35 items)
  - Plus additional 15 two-syllable words with different vowels (distractors)
  - Random order for each participant
  - Praat MFC (multiple forced-choice environment)
  - A short training set (6 items)



(a) (b)

Fig. 1: Evaluation of [ɪ] and [iː] vowels in experiment 1. The vowel is manipulated both in duration and formant values, Fratio stands for ratio on the range between formant values of natural [ɪ] and [iː]. Shades of grey represent mean values of evaluated vowel lengths from (a) 5 realisations of ASR, (b) 20 participants of the listening test (white = long, black = short).

Fratio=1, dur=90ms | Fratio=1, dur=195ms | Fratio=1, dur=300ms

Fratio=0.5, dur=90ms | Fratio=0.5, dur=195ms | Fratio=0.5, dur=300ms

Fratio=0, dur=90ms | Fratio=0, dur=195ms | Fratio=0, dur=300ms

Experiment 1

Statistic significance of duration and Fratio

- Mixed-effects model with logistic regression (binomial family for binary outcome)
- Both fixed effects were centred and standardised, replication was a random effect
- Random slopes included
- p-values obtained by likelihood ratio test

$$length \sim duration + Fratio + (1 + duration + Fratio | replication)$$

- ASR: Fratio and duration effects: $p < 0.001$
- ASR Subset ($duration \geq 160$ ms): Fratio $p < 0.001$, duration $p = 0.2554$
- Humans: both $p < 0.001$

Experiment 2 and 3

- [u] vs [uː], minimal pair: [kruciː] (meaning *cruel* in plural) and [kruːciː] (meaning *turkey* adjective)
- Manipulations of the 1st syllable (19 formant steps × 7 duration steps = 133 items)
  - [u] F1–F3 = 360, 906, 2774 Hz
  - [uː] F1–F3 = 288, 567, 2774 Hz
  - Duration 90 ms to 300 ms
  - ASR, 5 replications of random permutations
- Experiment 3: "zoom in" the transient area
  - Formants: 21 steps covering lower two thirds of Ex. 2
  - Durations: 125 ms – 230 ms in 9 steps

Fratio=1, dur=90ms | Fratio=1, dur=195ms | Fratio=1, dur=300ms

Fratio=0.5, dur=90ms | Fratio=0.5, dur=195ms | Fratio=0.5, dur=300ms

Fratio=0, dur=90ms | Fratio=0, dur=195ms | Fratio=0, dur=300ms

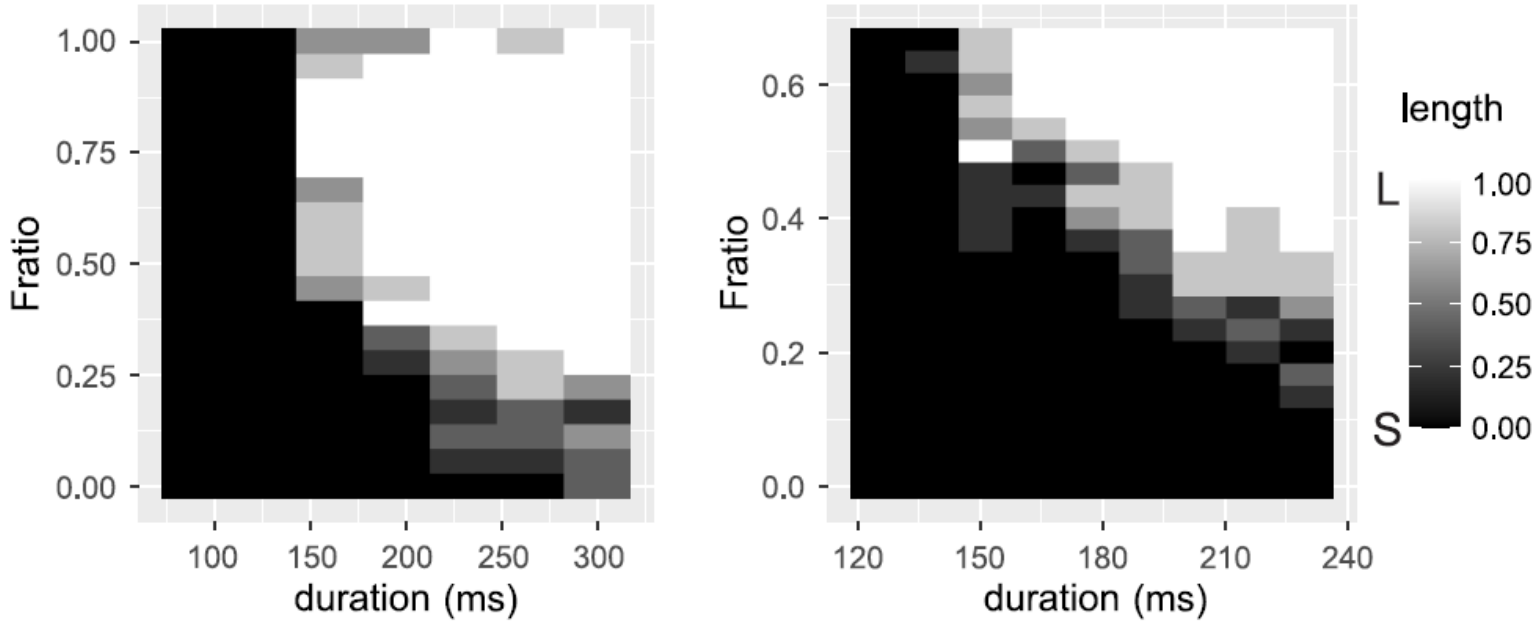


Fig. 2: Evaluation of [ʊ] and [uː] vowels. Fratio stands for ratio of the range between formant values of natural [ʊ] and [uː]. Shades of grey represent mean values of 5 evaluated vowel lengths by ASR in (a) experiment 2, (b) experiment 3 (white = long, black = short).

- Both $p < 0.001$

Fratio=1, dur=90ms | Fratio=1, dur=195ms | Fratio=1, dur=300ms

Fratio=0.5, dur=90ms | Fratio=0.5, dur=195ms | Fratio=0.5, dur=300ms

Fratio=0, dur=90ms | Fratio=0, dur=195ms | Fratio=0, dur=300ms

Conclusions

- ASR trained on an extensive sample of population reached results comparable with listening tests conducted on human subjects
- Advantages of ASR approach: unlimited number of items – parameters covered in much more detail, can be repeated many times with different settings
- Disadvantages of ASR: not possible to control subgroups (region, sex, age), more frequent errors in recognition of a word
- Combination with human perception experiments is recommended
- The results comply with recent study with one-syllable pseudo-words (Podlipský et. al 2019)
- Teachers of L2 Czech learners should be aware that perception of phonological high-vowels length is influenced by the tongue position

Thank you for your attention.