Perceived Length of Czech High Vowels in Relation to Formant Frequencies Evaluated by Automatic Speech Recognition

Tomáš Bořil(✉) and Jitka Veroňková

Faculty of Arts, Institute of Phonetics, Charles University, Nám. Jana Palacha 2,
Praha 1, Czech Republic
{tomas.boril,jitka.veronkova}@ff.cuni.cz

Abstract. Recent studies measured significant differences in formant values in the production of short and long high vowel pairs in the Czech language. Perceptional impacts of such findings were confirmed employing listening tests proving that a perceived vowel length is influenced by formant values related to a tongue position. Non-native speakers of Czech may experience difficulties in communication when they interchange the vowel length in words, which may lead to a completely different meaning of the message. This paper analyses perception of two-syllable words with manipulated duration and formant frequencies of high vowels i/iː or u/uː in the first syllable using automatic speech recognition (ASR) system. Such a procedure makes it possible to set a fine resolution in the range of examined factors. Our study confirms the formant values have a substantial impact on the perception of high vowels' length by ASR, comparable to mean values obtained from listening tests performed on a group of human participants.

Keywords: High Czech vowels · Vowel length · Vowel quality · Automatic speech recognition · Perception

1 Introduction

The acquisition of a vowel system is one of the key aspects of learning a second language (L2). Czech vowel system consists of five pairs of short and long monophthongs and three diphthongs [8,12,13]. Since the vowel length is phonologically distinctive, its improper interchange in L2 speakers' production may lead to a misunderstanding (e.g.., /kruːciː farmaːr̝i/ vs /kruciː farmaːr̝i/ (meaning *turkey farmers* vs *cruel farmers*).

This research was supported by the Czech Science Foundation project No. 18-18300S "Phonetic properties of Czech in non-native and native speakers' communication".

P. Sojka et al. (Eds.): TSD 2020, LNAI 12284, pp. 409–417, 2020.
https://doi.org/10.1007/978-3-030-58323-1_44

Differences between formant values (i.e., vowel quality correlating with a tongue setting in a vocal tract) of a short and a long vowel in a pair are traditionally described as insignificant (both in production and perception perspective) except for /i/ and /iː/ [5]. Later, a differentiation of short [ɪ] and long [iː] symbols in the international phonetic alphabet (IPA) was proposed [4].

In addition to statistical evaluation of production data of [ɪ] and [iː], [10] performed a perception analysis of manipulated items with a stimulus array covering the spectral and the durational span between both vowels in one syllable where both lengths create meaningful words with a comparable probability frequency. The study also found a significant difference in the perception of Bohemian (the western part of the Czech Republic) and Moravian (the eastern region of the Czech Republic) where Bohemians relied more on the spectrum, whereas Moravians relied more on the duration. Later, [14] found differences in pronunciation of [ɪ] and [iː] in the speech of Czech Radio newsreaders.

[12] focuses on the production of Czech in Bohemian and Moravian regions, and there is a clear trend of the [u] vs [uː] formant shift in the Bohemian subgroup in addition to the previously observed [ɪ] and [iː] relation. Spontaneous Czech speech was analysed in [7], the formant shifts between [ɪ] – [iː] and [u] – [uː] were measured, and also a promising difference between the [o] and [oː] formant positions appeared. [9] conducted a listening test with artificial one-syllable pseudowords containing manipulations of [ɪ] – [iː] and [u] – [uː] vowels analogous to [10] experiment. In both Czech high vowels, the quality (formant values) played a crucial role in a vowel length discrimination in the subgroup of listeners from the Bohemian region.

The main purpose of this paper is to compare the automatic speech recognition (ASR) of Czech high vowels' length with human perception in 3 experiments.

To emphasize the difference of qualities, we decided to use the [ʊ] IPA symbol for the short vowel and [uː] for the long vowel in the following text.

Experiment 1 examines perception of quantity (phonological length) of vowels [ɪ] and [iː] based on their quality (formant values in the spectrum). ASR evaluates items manipulated with a fine resolution in both duration and formant dimensions. A subset of this data set with a less detailed formant scale is also evaluated perceptually by human participants (Bohemian region) in a listening test. The question is whether ASR perceives the boundary between short and long vowels in a comparable manner and whether these results correspond to [10].

Experiments 2 and 3 analyse ASR behaviour on vowels [ʊ] and [uː] manipulated similarly. Experiment 3 focuses on the fine detail of the transition part found in experiment 2. The question is, whether the effect of formant values has an impact on the perceived length in compliance with the novel findings in [9], where artificial one-syllable pseudowords were tested by human participants in a listening test.

The ASR approach applied in this study may bring several advantages. The number of items in a listening test is naturally limited due to the requirement of keeping human participants entirely focused. For this reason, the number of tested factors and the resolution of coverage of their span have to be notably decreased in many experiments. The purpose of such experiments is to map a subjective perception of random individuals and then to estimate the mean value of the population. ASR systems are trained on a large sample of the population, and hence they also may provide evaluation similar to an average representative of the population. Such a procedure can be repeated many times with different settings and a large number of items, which would be impossible with human participants of listening tests.

2 Method

2.1 Experiment 1

For the first experiment, we created 147 manipulated items (21 formant steps and 7 duration steps) using Praat [2] and rPraat [3] as follows. A minimal pair consisting of two words [vɪrɪ] (meaning *viruses*) and [viːrɪ] (meaning *vortices*) was chosen to serve as boundaries lying on a diagonal of a two-dimensional duration–formant space to be explored. The advantage of the analysis of vowel in the first syllable is that it is not prone to phrase-final lengthening [15].

We recorded both words in a slow speech rate by an adult female speaker in a quiet low-reverb room (PCM uncompressed, the sample rate of 32 kHz, 16-bit depth). Estimated median values of formant frequencies F1–F4 of the target (first syllable) short [ɪ] were 405, 2295, 2866, and 4099 Hz and of the long [iː] were 305, 2700, 3000, and 4099 Hz (we rounded the fourth formant values in both vowels to the same number because instantaneous values reached a large variability around roughly the same values in both short and long vowels). For the manipulation purposes, we chose the record of [viːrɪ] as a basis because stimuli with shorter durations of the target vowel can be easily created by truncating the original long vowel.

The upper-part spectrum of the basis stimulus obtained by a high-pass Hann filter with a cut-off frequency of 4500 Hz was stored as a separate signal to be returned to manipulated signals at the final step of stimuli creation to obtain a more natural sound with a full range of the spectrum.

To obtain the source (excitation) signal and formant object, the basis stimulus was resampled to 16000 Hz and processed using the Burg method of linear predictive coding (LPC) with a prediction order of 15 (leading to max. 7 formant frequencies detected), 25 ms segmentation window length with 5 ms time step and pre-emphasis frequency of 50 Hz. Note: preliminary, prediction orders of 16 and 15 were examined in all experiments, the order of 16 in experiment 1 lead to an unnatural, artificial distortion at high frequencies; the order of 16 was necessary for experiments 2 and 3. To avoid random jumps in the formant object, formant trajectories were subsequently smoothed by a formant-tracking algorithm with 4 formant tracks.

In the next step, formant frequencies in the time interval of the first vowel duration were manipulated between the values of [ɪ] and [iː] in 21 linear steps. For each step, the manipulated sound was created by filtering the source (excitation) signal with the formant filter.

Finally, to create the whole set of target stimuli, each sound file was obtained by a concatenation of the first part of the original basis stimulus (until the first vowel), the shortened vowel from formant-manipulated signals with the upper-part spectra signal superposed, and the remaining part of the basis stimulus. The target vowel was shortened to durations in the range from 90 ms to 300 ms in 7 linear steps.

Automatic Speech Recognition. To evaluate manipulated stimuli by an automatic speech recognition system (ASR), we concatenated all stimuli in a random order into one long sound file. Each item was separated by a short pause and a neutral nonmanipulated word [vlakɪ] (meaning *trains*) by the same speaker to reduce possible interferences of two successive manipulated stimuli and also to clearly distinguish the boundaries of tested items in case the item was not recognized properly, e.g.., as two separate one-syllable words.

In total, we prepared five replications of the experiment, i.e., five different permutations of all manipulated items with different random order to avoid a possible effect of the order of stimuli.

To evaluate the concatenated sound file, we employed a commercial state-of-the-art ASR system Beey by NEWTON Technologies [6] set to the Czech language recognition and with additional text postprocessing switched off.

Although all nonmanipulated filler-words [vlakɪ] were recognized correctly, the ASR occasionally had problems with the recognition of manipulated items (probably due to their overall lower quality) and recognized them as a different word or a couple of two one-syllable words, e.g., [viːlɪ] (*fairies*), [bɪlɪ] (*they were*), [ɪ vɪ] (*also you*) or [bɪ jɪ] (*would her*). Not surprisingly, the consonants were affected, and the vowels remained either [ɪ] or [iː]. For this reason, we focused on the length of the first-syllable vowel [ɪ] or [iː] in such cases, ignoring mismatches in consonants.

For each item, the resulting score was calculated as a mean value of all five replications of the experiment.

Listening Test. To compare the results of ASR with human perception, we performed a listening test with 20 participants (native speakers of Czech, both male and female students, median age = 23 years) using comfortable headphones in a quiet room. To keep them focused throughout the test, we decided to select a subset of items only. The resolution of the vowel duration scale was kept the same, i.e., 7 linear steps between 90 ms and 300 ms. The resolution of formant transition was reduced to 5 discrete steps, resulting to 35 items. In addition to these "items-of-interest", other 15 two-syllable words with different vowels were included as distractors. Each of the total of 50 items in the test was initiated with a short desensitization beep sound.

The listening test was administrated using Praat multiple forced-choice (ExperimentMFC) environment [2]. After a short training set (6 items) to resolve possible problems and questions, the main test with 50 items in a random order for each listener was performed. Each item could be played three times at the most. The task was to click on a button with the word closest to the sound (both words with a short and a long vowel in the first syllable were offered). After the first 25 items, the participants were instructed to take a short break and listen to a song included in the test folder.

2.2 Experiment 2

In the second experiment, we created a set of stimuli focused on short [ʊ] and long [uː]. We recorded an adult male voice saying [krʊciː] (meaning *cruel* in plural) and [kruːciː] (meaning *turkey* adjective).

Estimated median values of formant frequencies F1 – F4 of the target (first syllable) short [ʊ] were 360, 906, 2774, and 3994 Hz, and of the long [uː] were 288, 567, 2774, and 3994 Hz (we rounded the third and the fourth formant values in both vowels to the same number because instantaneous values reached a large variability around roughly same values in both short and long vowels).

The process of manipulation was conducted in the same manner as in the experiment 1; the LPC prediction order was set to 16. The transition between two formant boundaries was divided into 19 linear steps. The duration of the vowel in the first syllable ranged from 90 ms to 300 ms in 7 linear steps.

This time, only the ASR task was performed with five random permutations of stimuli. Each item was concatenated with a nonmanipulated word [farmaːr̝ɪ] (meaning *farmers*), both variants creating a meaningful phrase with a similar probability frequency, i.e., the ASR should not prefer one variant at the expense of the other.

2.3 Experiment 3

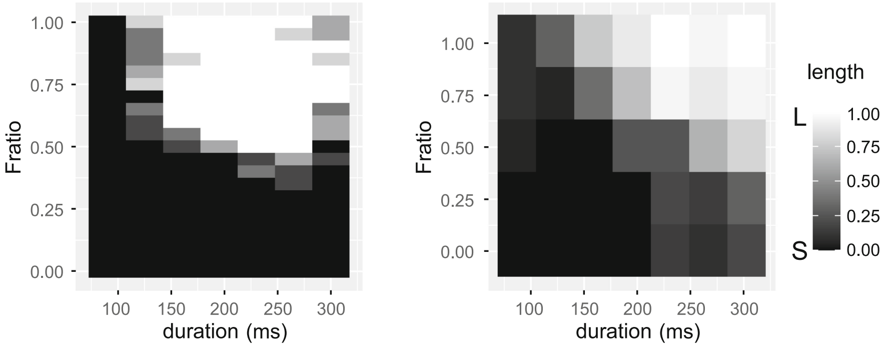
The third experiment continued with the same original records of experiment 2, but we aimed at the middle transient area. The formant axes were focused on the lower two thirds (as compared to experiment 2) with detailed 21 steps, and the duration focused on the middle part ranging from 125 ms to 230 ms in 9 detailed steps.

3 Results

We are aware of the fact our findings depend on a speech rate, a prosody, and an individual speaker's vocal space area; therefore we do not want to interpret our results as absolute values of boundaries between short and long vowel perception. Since this dependence can be a result of a complex combination of many factors, we do not even normalise duration and formant values because it could imply a universal rule. Rather than that, we focus on the shape of boundaries in the duration – formants relation which reflects the fact the vowel length perception is influenced by formant values, i.e., vowel quality.

3.1 Experiment 1

The results of ASR are depicted in Fig. 1a; a grey value of each rectangle represents a mean value of five replications of the experiment. Due to the statistic approach of ASR, some items were classified differently in some of the replications, which is mostly the case of items near the visible edge between short and long area.



(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).

The shortest items (duration of 90 ms) were identically identified as short vowel [ɪ]. All other items above the 90 ms duration were split into short [ɪ] and long [iː] with an almost horizontal boundary implying ASR used the vowel quality (i.e., spectrum) as the main cue to differentiate these two variants. This result complies with Bohemian Czech listeners in [9] (analysing artificial one-syllable words), although the ASR boundary seems slightly more horizontal.

We tested a statistic significance of duration and Fratio (a ratio on the span between typical formant values of the short and the long vowel) effects using mixed-effects models with logistic regression (binomial family for binary outcome) [1] in [11]. Both fixed effects (duration and Fratio) were centred and standardised, replication was a random effect. The model formula (including random slopes) is $length \sim duration + Fratio + (1 + duration + Fratio|replication)$, p-values were obtained by likelihood ratio tests of the full model with the effect against the model without the effect.

For both Fratio and duration effects, $p < 0.001$. We also passed a subset of data with a duration equal or larger than 160 ms, and for the Fratio, p-value remained <0.001; however, for the duration, $p = 0.2554$. This finding corresponds with 1a very well because vowel quality seems to be the main cue for longer durations.

Figure 1b represents the mean values of 20 participants of our listening test. These results are similar to ASR decision in Fig. 1a, although the Fratio scale is sampled in much fewer steps. However, for durations equal to or larger than 230 ms, some listeners evaluated items with Fratio = 0 (i.e., [ɪ]) as long. Statistical evaluation of both fixed effects was conducted analogously to the one with the ASR, subject (human participant) being a random effect. For both Fratio and duration, $p < 0.001$.

3.2 Experiment 2

Figure 2a represents results of ASR evaluating records with manipulated [ʊ]/[uː] vowels. For durations lower than or equal to 125 ms, all vowels were recognized as short despite the Fratio. For longer durations, the effect of the Fratio is visible. For both Fratio and duration, $p < 0.001$.

These ASR results are closely comparable to the relations observed in the listening tests of one-syllable pseudo-words in [9].

3.3 Experiment 3

The results of experiment 3 (i.e., detail zoom of the transition area of experiment 2) are depicted in Fig. 2b. The impact of vowel quality on recognized length is apparent and compatible with observations in [9]. For both Fratio and duration factors, $p < 0.001$.

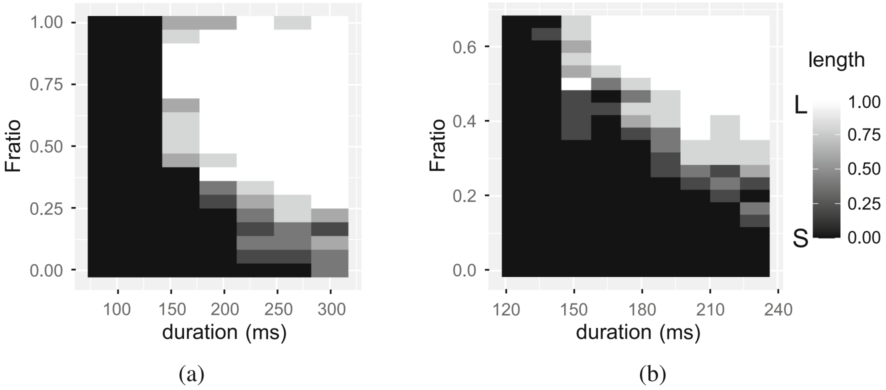


Fig. 2. Evaluation of [ʊ] and [uː] vowels. Fratio stands for ratio of the range between formant values of natural [ʊ] and [uː]. Shades of grey represent mean values of 5 evaluated vowel lengths by ASR in (a) experiment 2, (b) experiment 3 (white = long, black = short).

4 Conclusions

In the task of the evaluation of perceived phonological vowel length, ASR trained on an extensive sample of population reached results comparable with listening tests conducted on human subjects. The recent findings of the impact of vowel quality on perceived length of Czech high vowels in one-syllable pseudo-words [9] were confirmed on real two-syllable words in this paper.

Due to its phonological status in Czech, a mismatch in the vowel length could lead to misunderstandings and generally difficult communication, which is typical of foreign learners of the Czech language. Interestingly, based on our informal observation, we can say the vast majority of naïve L1 users of Czech language is not aware of these differences in quality of short and long pairs of high vowels. Teachers of L2 Czech learners, especially during the pronunciation training, should be aware of the fact that their perception of the phonological vowel length could be influenced not only by the vowel duration but also by the quality of short and long high vowel pairs.

The ASR technique may bring advantages in the process of evaluation in such a way that the count of items is not limited and the range of examined parameters can be covered in much more detail than in listening tests. On the other hand, a combination with these perception experiments is recommended as they may uncover additional effects such as region, sex or age of the listener.

References

1. Bates, D., Mächler, M., Bolker, B., Walker, S.: Fitting linear mixed-effects models using lme4. J. Stat. Softw. **67**(1), 1–48 (2015)
2. Boersma, P., Weenink, D.: Praat: doing phonetics by computer [Computer program]. Version 6.1.10 (2020). http://www.praat.org/
3. Bořil, T., Skarnitzl, R.: Tools rPraat and mPraat. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2016. LNCS (LNAI), vol. 9924, pp. 367–374. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-45510-5_42
4. Dankovičová, J.: Czech. J. Int. Phonetic Assoc. **27**(1–2), 77–80 (1997)
5. Hála, B.: Akustická podstata samohlásek. Česká akademie věd a umění (1941)
6. NEWTON Technologies: Beey [web-based platform]. Version 0.7.16.5 (2020). https://editor.beey.io
7. Paillereau, N., Chládková, K.: Spectral and temporal characteristics of Czech vowels in spontaneous speech. AUC PHILOLOGICA **2019**(2), 77–95 (2019)
8. Palková, Z.: Fonetika a fonologie češtiny: s obecným úvodem do problematiky oboru. Univerzita Karlova, vydavatelství Karolinum (1994)
9. Podlipský, V.J., Chládková, K., Šimáčková, Š.: Spectrum as a perceptual cue to vowel length in Czech, a quantity language. J. Acoust. Soc. Am. **146**(4), EL352–EL357 (2019). Acoustical Society of America
10. Podlipský, V.J., Skarnitzl, R., Volín, J.: High front vowels in Czech: a contrast in quantity or quality? In: Proceedings of Interspeech, vol. 2009, pp. 132–135 (2009)
11. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2020). https://www.R-project.org/

12. Šimáčková, Š., Podlipský, V.J., Chládková, K.: Czech spoken in Bohemia and Moravia. J. Int. Phonetic Assoc. **42**(2), 225–232 (2012)
13. Skarnitzl, R., Šturm, P., Volín, J.: Zvuková báze řečové komunikace: Fonetický a fonologický popis řeči. Univerzita Karlova v Praze, Nakladatelství Karolinum (2016)
14. Skarnitzl, R.: Dvojí i v české výslovnosti. Naše řeč **95**(3), 141–153 (2012)
15. Volín, J., Skarnitzl, R.: Temporal downtrends in Czech read speech. In: Proceedings of Interspeech, vol. 2007, pp. 442–445 (2007)