DURATION AS A CUE FOR PHONOLOGICAL VOICING CONTRAST IN WHISPERED CZECH

Michaela Svatošová, Tomáš Bořil

Institute of Phonetics, Faculty of Arts, Charles University, Prague, Czech Republic
michaela.svatosova@atarien.com, tomas.boril@ff.cuni.cz

ABSTRACT

In Czech, the phonological contrast of voicing is primarily realized by the presence or absence of the fundamental frequency. However, this main correlate is missing in whisper, because the vocal folds do not vibrate. The present study explored the acoustical and perceptual side of this phonological contrast. Firstly, it compared the duration of voicing counterparts of Czech plosives and fricatives embedded in modal and whispered pseudowords. The duration of voicing counterparts differed significantly, but the durational ratios were smaller in whisper than in modal phonation. Secondly, a perception experiment was created from the whispered stimuli, assessing the recognisability of whispered obstruents in phonetic context only. Listeners recognised many obstruents especially in the medial position, but substantial variability between individual voicing pairs was found.

Keywords: phonological voicing, whisper, Czech, perception, duration

1. INTRODUCTION

Whisper is used in various communicative situations with many functions [1]. In most cases, speakers want to be understood and the ongoing usage of whisper therefore suggests its (at least partial) intelligibility. Whisper is a type of phonation defined by the absence of vocal folds vibration (the only source signal is noise). However, in many languages, the presence or absence of voicing is exploited phonologically and some minimal pairs could become indistinguishable in whisper. The situational context of a given utterance provides syntactic, semantic and pragmatic cues that usually disqualify one member of the minimal pair. Nevertheless, the question of the relative importance of these cues remains.

Experiments that placed whispered voicing pairs in isolated syllables or short words (limiting the context to phonetic cues) have showed that listeners were able to discriminate voicing counterparts with an above chance success [2, 3, 4]. Other studies have investigated various acoustic cues that could differentiate voicing counterparts, suggesting duration of the target consonants as the most promising parameter [3, 4, 5, 6, 7, 8]. However, the majority of the above mentioned research concerned English in which phonological voicing strongly depends on aspiration (less affected in whisper). A deeper examination of the production and perception of voicing pairs in languages like Czech, which base this contrast on phonetic voicing [9], could be beneficial. Phonological voicing has been rarely studied for whispered Czech, but [10] have found differences in the duration of fricative voicing counterparts and the ability of speakers to distinguish them in minimal phonetic context.

The present study continues in this line of research, but extends the set of voicing pairs to fricatives and plosives. It addresses the perceptual aspect through the evaluation of listeners' ability to identify the voicing counterparts in whisper solely on the basis of phonetic cues (in nonsense pseudowords). From the perspective of production, it compares the duration (a possible correlate of phonological voicing) of voicing counterparts in both phonation types. The main research questions can be summarised as follows:

- Is the phonological contrast of voicing perceptually retained in whisper?
- Do the voicing counterparts differ in duration (in both phonation types)?

2. METHOD

2.1. Material

The study included 7 pairs of Czech obstruents with the same place of articulation, four plosive and three fricative pairs: /p b/, /t d/, /c ɟ/, /k g/, /f v/, /s z/, /ʃ ʒ/. These 14 phonemes were combined with the vowel /a/ into simple CV syllables, subsequently forming trisyllabic pseudowords adhering to Czech phonotactics [9]. Only consonants in the first and second syllable were analysed (the *initial* and *medial* position). Each phoneme appeared in ten different pseudowords for each position, creating a total of

140 unique stimuli (each containing an initial and medial target obstruent). These were complemented by 70 filler pseudowords. All non-analysed syllables featured a wider range of phonemes in order to avoid monotony. Each combination of initial and medial target obstruent repeated maximally three times and the stimuli were controlled for similarity with real Czech words. A sample of target and filler pseudowords is presented in Table 1.

Target stimuli			Fillers
/**t**a**ʃ**amu/	/**z**a**v**ako/	/**f**a**d**aju/	/jovukɪ/
/**k**a**t**apɪ/	/**p**a**v**alɛ/	/**ʒ**a**ɟ**afo/	/lɛxara/

Table 1: Examples of target and filler pseudowords. Target phonemes are in bold.

2.2. Speakers and recording

All stimuli were randomly divided into 5 blocks (the assignment was done separately for each phonation). The pseudowords were written in normal Czech ortography and they were placed in lines of six (separated by commas) with the first and last word randomly chosen from the set of fillers. The stimuli were recorded in a sound-treated studio by 10 native speakers of Czech (5 men, 18–26 years old) with no speech disorders. They did not receive any reward. Each of them had a unique order of blocks, but modal and whispered blocks always alternated. The speakers were instructed to read with normal effort and at moderate tempo, separating each word with a small pause. Stimuli produced in an incorrect phonation or with hesitation were repeated at the end of the session. The recordings were saved in uncompressed PCM format with a sampling frequency of 44.1 kHz and with 24-bit quantisation.

2.3. Acoustical analysis

All recordings were manually annotated in Praat [11] following the segmentation rules described in [12]. Whenever the nature of whispered speech prevented the use of the typical cues, the placement of boundaries was guided mainly by the changes of amplitude in the oscillogram. The analysis of the temporal data was performed in R using the packages *rPraat*, *tidyverse* and *boot* [13, 14, 15, 16]. In order to compare data from speakers with different articulation rates, the duration of each phone was multiplied by the mean articulation rate of the given speaker and divided by 10. The normalised values thus corresponded to the articulation rate of 10 phones/second. Only phones in the first two syllables were used for the calculation of the mean articulation rate, because the final syllables could have been affected by final lengthening and determining their precise boundaries was problematic. The duration of plosives was not measured in the initial position.

The effect of voicing on normalised duration was statistically evaluated with linear mixed effects models using the package *lme4* [17] in R. The fixed factors were (phonological) voicing, voicing pair, sex and position; random factors were speaker (with random slopes for voicing) and pseudoword. The significance of the factor of voicing and its interactions with other fixed factors was determined by *Likelihood Ratio Tests*.

2.4. Perception experiment

The perception experiment contained stimuli from 4 speakers (2 men). There were 28 whispered pseudowords (representing each target obstruent in both positions) selected from each of them. The stimuli were chosen to have duration values close to the mean values of the given target phoneme and speaker. Each pseudoword occured only once in the whole experiment. The participants (29 native speakers of Czech, 9 men, the interquartile range of age 22–53 years) were presented with a word in which the target phoneme was replaced with an underscore. They could replay the recording once. They chose the missing phoneme from a list that contained 19 consonants that occured in the stimuli. The target stimuli were complemented by 32 fillers (8 from each speaker), asking for non-target consonants. The stimuli followed in a random order. The participants practiced the operation of the experiment with 8 training stimuli. The perception experiment was performed online using Psytoolkit [18, 19] and it took 10–20 minutes. The participants could take two breaks and were instructed to wear headphones and to do the experiment in a silent setting. They did not receive any reward.

Although listeners could choose from a list of 19 consonants, the answers that incorrectly identified features other than voicing amounted to only 2% of all responses and were excluded from further analyses. An identification index was calculated for each listener and target obstruent (e.g. 0.75 for medial /s/, indicating that the given listener recognised 3 of 4 occurences of that obstruent). These indices served for the statistical analysis of the mean identification score of each group with the binomial test using the package *Hmisc* [20] in R. The result was considered significant if the error bar did not include the value 0.5 (chance level).

3. RESULTS

3.1. Duration of target obstruents

Table 2 provides the durational data of the target obstruents in modal phonation. The effect of voicing on normalised duration was statistically significant ($\alpha = 0.05$, $p < 0.001$), as were the interactions of voicing with the fixed factors of voicing pair ($p < 0.001$), position ($p < 0.001$) and sex ($p = 0.0031$).

voicing pair	position	difference (ms)	ratio
/p b/	medial	31.0	1.36
/t d/	medial	42.9	1.63
/c ɟ/	medial	34.9	1.40
/k g/	medial	34.4	1.42
/f v/	initial	53.8	1.55
	medial	54.5	1.72
/s z/	initial	38.4	1.33
	medial	58.3	1.62
/ʃ ʒ/	initial	37.1	1.30
	medial	50.7	1.52

Table 2: Differences and ratios of mean normalised duration of phonologically voiceless to voiced obstruents in modal phonation.

The same measurements were taken for whispered stimuli, as shown in Table 3. The effect of voicing on normalised duration was also statistically significant ($\alpha = 0.05$, $p < 0.001$), as were the interactions with the fixed factors of voicing pair, position and sex ($p < 0.001$ for all interactions). Higher values of voiceless counterparts were present among all voicing pairs, with mean differences in lower tens of milliseconds.

voicing pair	position	difference (ms)	ratio
/p b/	medial	23.6	1.25
/t d/	medial	37.9	1.50
/c ɟ/	medial	25.1	1.27
/k g/	medial	17.8	1.20
/f v/	initial	60.0	1.88
	medial	41.2	1.52
/s z/	initial	36.2	1.36
	medial	34.5	1.35
/ʃ ʒ/	initial	28.9	1.27
	medial	31.1	1.33

Table 3: Differences and ratios of mean normalised duration of phonologically voiceless to voiced obstruents in whispered phonation.

Since longer duration of voiceless obstruents was found in both phonation types, Figure 1 compares the mean voiceless to voiced ratios of normalised duration in both. In medial position, the durational contrast was significantly smaller in whisper for four voicing pairs and the same tendency was present in the other three pairs. On the other hand, modal initial fricatives had similar ratios as their whispered variants. The initial /f v/ showed great variability in duration (also enhanced by the difficulties in delimiting their boundaries), which reduces the reliability of the respective results.

3.2. Perception experiment

Listeners have successfully identified 78% of all target obstruents (success rates ranged between 61–87% for individual listeners). Figure 2 provides the mean identification scores for each target obstruent in both positions. There were apparent differences – while the pairs /t d/ and /f v/ were recognised in both positions, none of the palatal plosives reached statistical significance. The group of initial voiced obstruents tended to have the lowest scores.

4. DISCUSSION

The modal stimuli have mostly confirmed the durational differences reported by [10, 21], although the ratios for fricatives were 20–30 percentage points lower (but in the same order of voicing pairs). The results for whispered fricatives were also in line with [10]. Shorter duration of voiced obstruents in modal phonation is caused by their complex articulation, which includes conflicting aerodynamic conditions [22]. However, the same constraints do not apply for whisper and therefore the observed differences (although smaller than in modal phonation) have to be explained differently.

The perception experiment showed that listeners were quite successful in discriminating the voicing counterparts solely on the basis of phonetic cues. Interestingly, the recognisability of individual obstruents can be linked to the durational ratios of the respective voicing pairs – the pairs /t d/ and /f v/ had the highest ratios (above 1.5), while the bilabial and palatal plosives had the lowest ones (up to 1.3). The importance of duration would also explain another pattern in the perception data, namely the higher recognisability of obstruents in medial position, since it provides more obvious boundaries for the target phones. It is also noteworthy that the listeners did not choose voiceless obstruents more often than voiced ones. On the contrary, they balanced their responses despite the absence of voicing and the bias created by Czech phonological inventory (voiceless obstruents are more frequent than voiced ones, except for the pair /f v/ [23]).

However, the high identification scores of

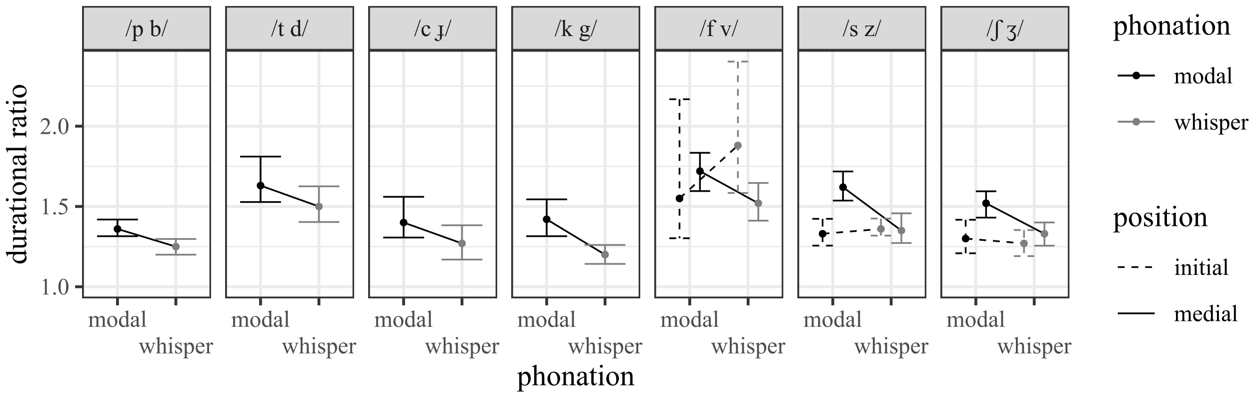

Figure 1: The ratios of the mean normalised durations of all target obstruents. The error bars indicate the 95% confidence intervals (with the Bonferroni correction for $n = 10$) calculated with the bootstrap method.

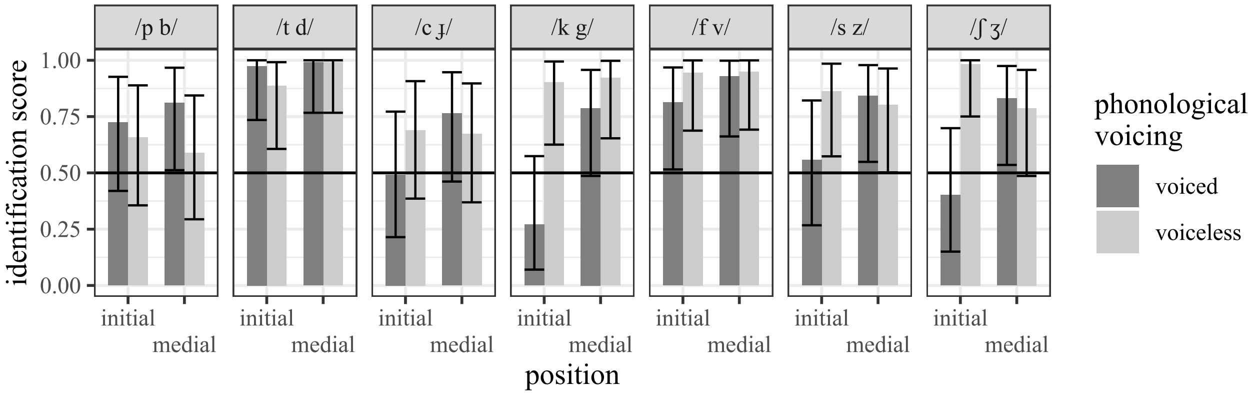

Figure 2: The mean identification scores for each target obstruent in both positions. The error bars indicate the 95% confidence intervals (with the Bonferroni correction for $n = 28$).

phonemes /t d/ and /f v/ are probably caused by multiple factors. Although they phonologically form pairs, their place and/or manner of articulation is considerably different. This is most prominent in the first pair, which consists of a dento-alveolar [t] that contrasts with a (post-)alveolar [d] (often realised as a tap [ɾ]) [24]. Similarly, the Czech /v/ acoustically resembles an approximant rather than a fricative, especially in intervocalic position [25]. Since the features of place and manner are relatively well preserved in whisper (as supported by [3] and the minor amount of other-than-voicing errors in the present perception experiment), they might have raised the identification scores of these two pairs. Moreover, both taps and approximants are associated with shorter durations, which further enhances the durational difference.

It seems that the phonological voicing contrast is partially preserved in whisper, especially in medial position. The voiced obstruents were found to be significantly shorter than their voiceless counterparts. This result supports the *redundant cue hypothesis* as described by van de Velde and van Heuven [8], because the durational ratios in whisper were less pronounced than in modal phonation and speakers were probably not trying to enhance the differences between voicing counterparts in order to compensate for the absent voicing.

Further research is needed to overcome the limitations of the present study, which focused on obstruents in isolated pseudowords with simple syllabic structure. Moreover, the material consisted of read speech and the results cannot be generalised to spontaneous interactions. The production and perception part suggested a connection between durational ratios and recognisability of voicing pairs in whisper, but only indirectly, since the stimuli were not controlled for other parameters. It would be desirable to perform perception experiments with manipulated stimuli differing only in duration.

5. REFERENCES

[1] J. Cirillo, "Communication by unvoiced speech: the role of whispering," *Annals of the Brazilian Academy of Sciences*, vol. 76, no. 2, pp. 413–423, 2004.

[2] G. L. Dannenbring, "Perceptual Discrimination of Whispered Phoneme Pairs," *Perceptual and Motor Skills*, vol. 51, no. 3, pp. 979–985, 1980.

[3] V. C. Tartter, "What's in a whisper?" *The Journal of the Acoustical Society of America*, vol. 86, no. 5, pp. 1678–1683, 1989.

[4] T. Mills, "Cues to voicing contrasts in whispered Scottish obstruents," Master's thesis, University of Edinburgh, 2003.

[5] M. F. Schwartz, "Bilabial Closure Durations for /p/, /b/, and /m/ in Voiced and Whispered Vowel Environments," *The Journal of the Acoustical Society of America*, vol. 51, pp. 2025–2029, 1972.

[6] M. Parnell, J. D. Amerman, and G. B. Wells, "Closure and constriction duration for alveolar consonants during voiced and whispered speaking conditions," *The Journal of the Acoustical Society of America*, vol. 61, no. 2, pp. 612–613, 1977.

[7] S. T. Jovičić and Z. Šarić, "Acoustic Analysis of Consonants in Whispered Speech," *Journal of Voice*, vol. 22, no. 3, pp. 263–274, 2008.

[8] D. J. van de Velde and V. J. J. P. van Heuven, "Compensatory Strategies for Voicing of Initial and Medial Plosives and Fricatives in Whispered Speech in Dutch," in *Proceedings of the International Congress of Phonetic Sciences*, vol. 17, 2011, pp. 2058–2061.

[9] Š. Šimáčková, V. J. Podlipský, and K. Chládková, "Czech spoken in Bohemia and Moravia," *Journal of the International Phonetic Association*, vol. 42, no. 2, pp. 225–232, 2012.

[10] P. Machač and P. Šturm, "The phonological contrast of voicing in whispered Czech and its phonetic correlates – a preliminary study." in *20th Czech-German Workshop – Speech Processing*, R. Vích, Ed., 2010, pp. 34–43.

[11] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," 2020. [Online]. Available: https://www.praat.org/

[12] P. Machač and R. Skarnitzl, *Fonetická segmentace hlásek*. Praha: Epocha, 2009.

[13] R Core Team, "R: A Language and Environment for Statistical Computing," 2019. [Online]. Available: https://www.R-project.org/

[14] T. Bořil and R. Skarnitzl, "Tools rPraat and mPraat," in *Text, Speech, and Dialogue: 19th International Conference, TSD 2016, Brno, Czech Republic, September 12-16, 2016, Proceedings*, P. Sojka, A. Horák, I. Kopeček, and K. Pala, Eds. Springer International Publishing, 2016, pp. 367–374.

[15] H. Wickham, M. Averick, J. Bryan, W. Chang, L. D. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. L. Pedersen, E. Miller, S. M. Bache, K. Müller, J. Ooms, D. Robinson, D. P. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, and H. Yutani, "Welcome to the Tidyverse," *Journal of Open Source Software*, vol. 4, no. 43, pp. 1686–1691, 2019.

[16] A. Canty and B. D. Ripley, *boot: Bootstrap R (S-Plus) Functions*, 2021, R package version 1.3-28.

[17] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[18] G. Stoet, "PsyToolkit: A software package for programming psychological experiments using Linux," *Behavior Research Methods*, vol. 42, no. 4, pp. 1096–1104, 2010.

[19] ——, "PsyToolkit: A Novel Web-Based Method for Running Online Questionnaires and Reaction-Time Experiments," *Teaching of Psychology*, vol. 44, no. 1, pp. 24–31, 2017.

[20] F. E. Harrell, "Hmisc: Harrell Miscellaneous," 2021. [Online]. Available: https://CRAN.R-project.org/package=Hmisc

[21] P. Machač, "Temporální a spektrální struktura českých explozív," Rigorózní práce, Univerzita Karlova, Filozofická fakulta, Ústav obecné lingvistiky, Praha, 2006.

[22] J. J. Ohala, "Aerodynamics of phonology," in *Proc. 4th Seoul International Conference on Linguistics*, 1997, pp. 92–97.

[23] A. Bičan, "Phonological Corpus of Czech," 2016. [Online]. Available: https://ujc.avcr.cz/phword

[24] R. Skarnitzl, "Asymmetry in the Czech Alveolar Stops: An EPG Study," *AUC Philologica 1/2014, Phonetica Pragensia XIII*, vol. 2014, no. 1, pp. 101–112, 2014.

[25] R. Skarnitzl and J. Volín, "Czech Voiced Labiodental Continuant Discrimination from Basic Acoustic Data," in *Proceedings of Interspeech 2005*, 2005, pp. 2921–2924.