ACTA UNIVERSITATIS CAROLINAE
PHILOLOGICA 3/2017

ACTA UNIVERSITATIS CAROLINAE

PHILOLOGICA 3/2017

Editors
RADEK SKARNITZL and JAN VOLÍN

CHARLES UNIVERSITY
KAROLINUM PRESS
2017

Editors: doc. Mgr. Radek Skarnitzl, Ph.D.
doc. PhDr. Jan Volín, Ph.D.

http://www.karolinum.cz/journals/philologica

ISSN 0567-8269 (Print)
ISSN 2464-6830 (Online)

CONTENTS

FUNDAMENTAL FREQUENCY STATISTICS FOR MALE SPEAKERS OF COMMON CZECH

RADEK SKARNITZL AND JITKA VAŇKOVÁ

ABSTRACT

In speaker identification, a forensic phonetician's task often involves comparing voices of two or more speakers and assessing their similarity, but also their typicality. For the latter, it is necessary to have background information about the relevant speaker population. This paper introduces the database of Common Czech which was compiled as a reference database, and presents the first set of compiled statistics pertaining to fundamental frequency (F0). The population statistics are computed from a reading task and spontaneous speech. The results confirm the superiority of F0 baseline over mean or median values when assessing typicality and demonstrate, in many speakers, a narrower intonation range in spontaneous speech than in reading. The role of F0 in speaker comparison is also discussed.

Key words: fundamental frequency, forensic phonetics, speaker identification, Czech

1. Introduction

Voice comparison is probably the most frequent task within the large domain of speaker identification (Jessen, 2012; Skarnitzl, 2014). This task consists in comparing, by means of detailed acoustic and auditory analyses (Nolan, 1999), the voice in the unknown recording (originating, for instance, from an anonymous telephone call) with that in the known recording (obtained typically during a police interrogation). The determination of identity or non-identity of the voices depends on the extent to which various acoustic features and phenomena determined by listening (see, e.g., Hollien & Hollien, 1995; Hollien, 2002: 78ff. for analytic approaches to auditory speaker identification) are similar or different.

However, apart from looking for *similarities* and *differences* between the voices under comparison, it is necessary to take into consideration the *typicality* of the observed values (Jessen, 2012: 40ff.). It is clear, for example, that a difference between two voices of 10 Hz in the mean F2 frequency of the Czech vowel [uː] has a different explanatory power when

https://doi.org/10.14712/24646830.2017.29

the means are 780 and 790 Hz, and when they are 1280 and 1290 Hz. While the similarity is practically meaningless in the former case, as this is very close to the mean F2 value of the Czech [uː] (Skarnitzl & Volín, 2012 report 770 Hz for young Czech male speakers), the latter situation would be considerably more interesting in forensic phonetic casework. Such a high frequency of F2 in the [uː] vowel indicates substantial – and untypical – fronting and/or de-labialization. If combined with other acoustic and auditory features, a forensic phonetician's conclusion may be that it is much more likely for such similarities to arise when the voices under comparison originate from the same speaker than if they were to originate from different speakers.

In order to be able to make judgements about the typicality of observed similarities and differences in the measured values, these values must be related to the statistics of the relevant population. In other words, we need to know the patterning of each feature in the population: its mean or median value as a representative of the central tendency (i.e., the 770-Hz value of F2 for the [uː] vowel; see Volín, 2007a: section 3.3.1), standard deviation (SD) or other measures of the feature's variability (*ibid:* section 3.3.2), and possibly a more detailed expression of the shape of its distribution. Only then can we determine whether the observed similarities may speak for the hypothesis of identity of the voices.

Given the importance of population data for forensic practice, it should not be surprising that recording databases of target populations and compiling population statistics of various acoustic features rank among the most important tasks of forensic phonetic research. While historically, to the best of our knowledge, some of the first reference data were produced for German (Jessen, Köster & Gfroerer, 2005 for fundamental frequency; Jessen, 2007 for articulation rate), it is currently Standard Southern British English (SSBE) where the greatest advances are being made. The DyViS database (Dynamic Variability in Speech; Nolan, McDougall, de Jong & Hudson, 2009) features recordings of 100 young male speakers in various forensically relevant speaking tasks. Population statistics have so far been presented for fundamental frequency (Hudson, de Jong, McDougall, Harrison & Nolan, 2007), and less comprehensive data also exist for disfluency features (McDougall, Duckworth & Hudson, 2015). The comparative lack of more available population data serves to show how demanding the task of producing population data is – the requirements in terms of personnel and time are tremendous, and they are far from negligible when it comes to technical and phonetic expertise.

The term "relevant" or "target" population has been used above, without going into much detail. However, it is clear that the decision as to what is a relevant population in a specific case is by no means trivial (Morrison & Ochoa, 2012). In fact, there is an inherent paradox involved in this decision (Hughes & Foulkes, 2015): we are trying to establish the population (speech community) relevant for the given speaker without knowing the speaker's identity. Forensically applicable databases typically feature male speakers of one variety (determined regionally or socially), especially in languages with high regional variability. The question is far less straightforward when it comes to the age of the speakers. The above-mentioned DyViS database only includes speakers aged 18 to 25 years, presumably in part due to the relative accessibility of this age group at universities, but also due to the phonetic changes which SSBE is undergoing (de Jong, McDougall & Nolan, 2007). On the other hand, the German database included speakers of a rather wide age range, between 21 and 63 years.

It should be pointed out that a database of the relevant population is necessary not only in phonetic speaker identification, but also in automatic speaker identification, a factor which often seems to be overlooked in casework (Svobodová, 2014, personal communication). Automatic approaches typically exploit GMM-UBM, where the UBM stands for a universal background model; this represents the distributions of feature vectors of a general population of speakers (Reynolds, Quatieri & Dunn, 2000; Reynolds & Campbell, 2008), and it is clear that this population should correspond to some extent to the speakers under investigation.

The objective of this paper is twofold: we will report on the compilation of a new forensically applicable database of the Czech language, and we will present the first population statistics; as in most other languages, this will involve population data of fundamental frequency (F0). F0 distribution is a suitable parameter when beginning to create population statistics, because it belongs to the most commonly used parameters in forensic speaker comparison (Jessen, 2012: 67) and also thanks to the ease of its extraction.

2. Database of Common Czech

Until recently, there have been no forensically applicable population statistics for the Czech language. Only a modest reference set of manually measured formant values had been compiled for a small number of 27 male speakers by Skarnitzl & Volín (2012); however, this database, besides being limited in terms of the number of speakers, is also based only on a read text.

The choice of the variety for our database – Common Czech – was relatively straightforward. Common Czech is a supraregional, non-standard variety of Czech commonly used in everyday communications. It is often referred to as an interdialect which, in addition, has been affecting the standard variety (Krčmová, 2005; Chromý, 2014). Since Common Czech is used habitually for casual oral communication with family, friends or acquaintances, but it can also be encountered in more formal situations, it is the ideal variety on which population statistics of Czech should be based.

The recordings for the database were acquired during 2015. The database includes 100 male speakers aged between 19 and 50 (mean age: 25.6 years, SD: 6.7 years); this is the population which is most relevant in forensic phonetic contexts. Each person was recorded by a friend or an acquaintance to facilitate natural speech production. The recordings were obtained in quiet environments via a professional portable recorder Edirol HR-09 in a WAV format with 48-kHz sampling frequency. In order to gain a representative sample of the target speakers' speech performance, the recorded material covered several speaking styles (the stylistic differentiation follows the structure of the VASST corpus, which features predominantly older speakers from various regions of the Czech Republic). Every speaker completed six speaking tasks which lasted between 45 and 60 minutes in total:

1) a structured interview which comprised a pre-defined set of general questions (e.g., "What kind of music do you like?", "At which occasions do you listen to it?", "Do you like sports?"), as well as demographically oriented questions (e.g., "Where did you spend your childhood?", "Have you moved a lot?", "What foreign languages do you speak?")

2) a free, spontaneous interview taking approximately 25 minutes, in which the target speaker was encouraged to talk freely about any topic; the experimenters had a set of questions to get the speaker talking, but the speakers were free to choose any topic they felt like talking about, so that their speech was as spontaneous as possible
3) reading a phonetically rich text of 150 words which includes all phonemes and their context-dependent variants; the speakers were given time to familiarize themselves with the text
4) a picture description task lasting approximately 6 minutes
5) reading specific phrases and sentences
6) a disguise task in which the speakers were asked to disguise their voice so that it is not recognizable; they were given time to decide what kind of disguise they would use, and then they read a text which contained similar word sequences as task 3 (see also Růžičková & Skarnitzl, 2017)

As of March 2017, the recordings have been cut into the six parts and downsampled to 32 kHz. Recordings of the read tasks (numbers 3 and 6 above) have been segmented using the Prague Labeller, an HMM-based forced alignment tool (Pollák, Volín & Skarnitzl, 2007). In addition, the text of approximately 20 of the free interviews (task 2 above) has been transcribed, and it is hoped that this work will continue. While none of the tasks in the database were transmitted over a mobile phone, we plan to simulate mobile transmission by compressing the signal using the GSM AMR codec (3GPP, 2012), the most widespread codec in mobile telephones (Vaňková & Bořil, 2014).

3. Population statistics of fundamental frequency

Fundamental frequency (F0) is the acoustic correlate of the frequency of vocal fold vibration, and as we have mentioned above, F0 population statistics belong to those which are available most frequently for a given language. There are several ways of capturing the speaker-specific behaviour of F0 (see Rose, 2002: Chapter 8, Jessen, 2012: Chapter 3, or Skarnitzl & Hývlová, 2014 for more details and also for methodological issues related to using F0 in forensic phonetic contexts). It is beneficial to be able to determine a central value of F0 for a given speaker, one which characterizes his or her modal phonation behaviour; such a value has been called the *speaking fundamental frequency* (SFF). Researchers have most frequently used the arithmetic mean or median value to refer to SFF. More recently, the so-called F0 baseline has been proposed by Lindh & Eriksson (2007) as a speaker's neutral, carrier F0. This value, which is supposed to be most robust vis-à-vis various technical and behavioural distortions, is computed as the 7.64th percentile of a speaker's F0 dataset. All these three expressions of SFF will be provided for Common Czech, along with measures of variability.

3.1 Method

Since both reading and spontaneous speech are typically used in police interrogations to obtain speech samples for comparison, F0 was examined in two of the speech styles mentioned above: in the spontaneous interview (task 2; ca. one minute was selected from the central portion of the recording) and in ordinary reading (task 3, which typically lasted between 60 and 75 seconds). One minute of speech has been shown to be sufficient to determine a speaker's SFF (e.g., Nolan, 1983: 123; Volín, 2007b).

F0 values were extracted automatically using autocorrelation in Praat (Boersma & Weenink, 2016), with the interval of 10 msec and extraction range being 60–350 Hz; all other settings were kept at their default values. The extracted values were not corrected manually, partly due to the vast amount of data, and also, more importantly, because manual corrections are out of the question in actual forensic phonetic casework, which tends to be severely time-constrained.

The raw F0 data were processed in R (R Core Team, 2015) and visualized using the package *ggplot2* (Wickham, 2009).

3.2 Results and discussion

Table 1 shows the overall mean values of three SFF indicators in spontaneous male Czech speech and in reading. As we can see, the mean and median values are lower in spontaneous speech than in reading, and this difference is highly significant (t-test for repeated measures for mean values: $t(99) = 11.4$, $p < 0.0001$; for median values: $t(99) = 12.9$, $p < 0.0001$). This result is in agreement with studies of F0 in English (Hirson, French & Howard, 1995; Hollien, Hollien & de Jong, 1997); interestingly, an opposite but insignificant tendency has been observed in German (Jessen et al., 2005).

Perhaps a more interesting result, however, is the fact that the mean value of F0 baseline changes the least between the spontaneous interview and the reading task – the difference is not significant: $t(99) = 1.5$, $p > 0.1$. This lends support to Lindh and Eriksson's (2007) claim that this empirically derived indicator of F0 central tendency is more robust to changes in speaking style than the mean or median values.

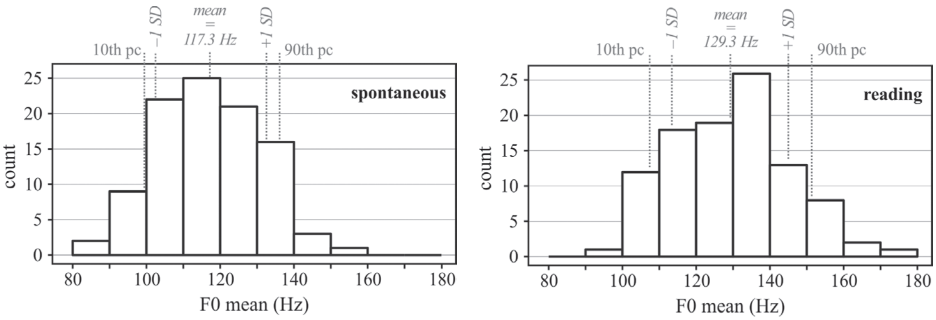
Table 1. Indicators of speaking fundamental frequency and of F0 variability among 100 male Czech speakers in spontaneous and read speech.

	spontaneous	reading
mean (Hz)	117.3	129.3
median (Hz)	113.1	125.8
baseline (Hz)	95.4	96.8
SD (Hz)	15.0	15.9
varco (%)	12.8	12.3

The table also displays two variability measures, standard deviation (the most frequently given measure) and coefficient of variation, or *varco*, which relates SD to the mean and is thus more straightforward (although the benefit is marginal with the mean F0 values lying close to 100 Hz). These figures for all the speakers in our corpus indicate that the overall variability does not change between spontaneous and read speech.

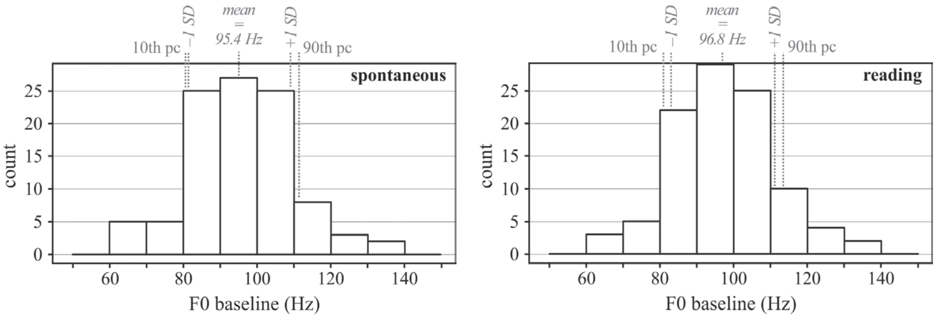
Next, we will consider the distribution of F0 data in more detail. In order to be able to compare our results with studies conducted on other languages, we will use the mean value, but let us repeat our conviction that the baseline has a lot of potential in F0 analysis; that is why baseline distributions will be described as well. Figure 1 shows the distribution of the mean values of the 100 speakers in 10-Hz intervals, for both spontaneous and read speech. Figure 2 shows the histograms for baseline values. The overall means (*cf.* Table 1), as well as the ± 1 SD and 10th–90th percentile ranges from the mean are indicated in the figures.

Figure 1. Histograms of the F0 mean values among 100 male Czech speakers in spontaneous speech (left) and in reading (right); the overall mean and the range ± 1 SD from the mean and the 10th–90th percentile (pc) range are indicated in grey.



We can infer from the histogram of mean F0 values in spontaneous speech (top part of Fig. 1) that mean F0 is situated between 110 and 120 Hz in 25 speakers (i.e., 25 percent of all speakers) and that more than two thirds (68%) of mean values are located in the 100–130 Hz range. Similarly, we can see from Fig. 2 that F0 baseline is located within the 80–110 Hz range in more than three quarters of the speakers in spontaneous speech, as well as in reading. It is these distributions which will allow forensic practitioners to judge the relevance of the similarities or differences between analyzed voices: if SFF values of two voices fall within the same range as most speakers in the population, the finding cannot contribute to the hypothesis of identity or non-identity in any way; if, on the other hand, the SFF of one or more of the voices under comparison falls below or above the most typical 30Hz interval, the information may be relevant.

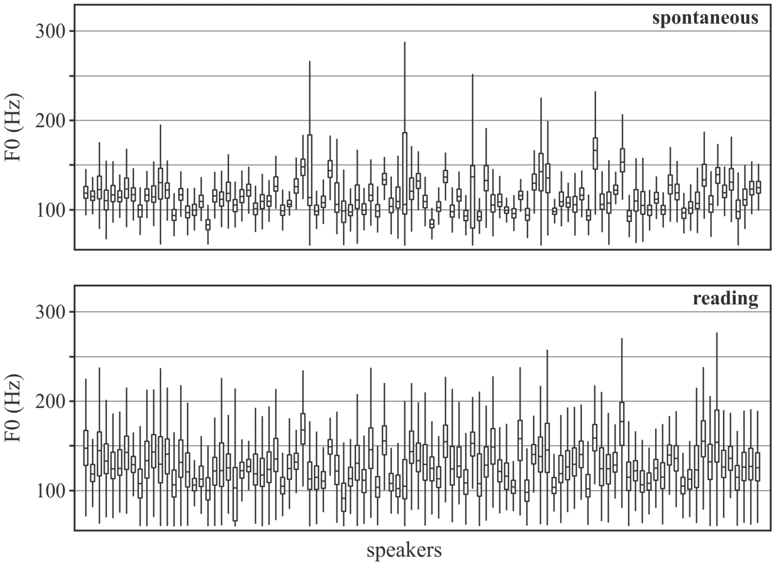
Figure 2. Histograms of F0 baseline values among 100 male Czech speakers in spontaneous speech (left) and in reading (right); the overall mean and the range ± 1 SD from the mean and the 10th–90th percentile (pc) range are indicated in grey.



The objective of the following analysis is to provide at least a small glimpse into the variability of individual speakers. This was partly motivated by our preliminary results based on 26 speakers from the current corpus which showed markedly lower F0 variability in spontaneous speech than in reading (Skarnitzl & Vaňková, 2015). The overall difference in the entire dataset of 100 speakers, however, is negligible, with the mean value of the coefficient of variation being 21.1% in spontaneous and 21.5% in read speech (*t*-test for repeated measures: $t(99) = 0.6$, $p > 0.5$).

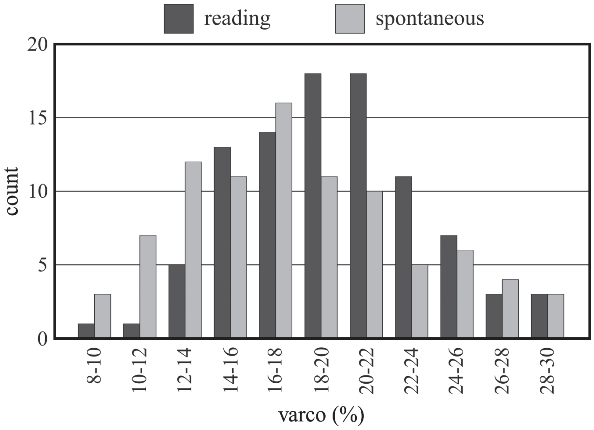
Figure 3 provides a more detailed look at the behaviour of individual speakers, and we can see that in many speakers the variability does appear to be lower in spontaneous speech. Some speakers, on the other hand, produced an extremely wide range of F0 values, so that the differences therefore seem to average themselves out. It must be noted that, first, outlier values are not plotted in Figure 3 for the sake of clarity and second, that the F0 tracking has not undergone manual correction, as noted above, and the F0 extraction error rate may be greater for some speakers than others. A cursory inspection of the recordings and F0 tracks indicates that this does seem to be the case with the highest values depicted in Figure 3.

Figure 3. Variability of F0 in the individual speakers in spontaneous speech (top) and in reading (bottom). The boxes indicate quartile ranges, whiskers denote ±1.5 IQR from the quartiles; outlier values are not plotted.



We can also analyze the melodic variability by comparing the coefficients of variation in reading and in spontaneous speech, as shown in Figure 4. We are interested especially in the lower end of the scale, where we can see that, indeed, a greater number of speakers manifests low *varco* values in spontaneous speech than in reading. We may thus tentatively conclude that quite a lot of speakers in our database manifest a narrower intonation range in spontaneous speech than in reading. Flatter intonation in many speakers' spontaneous interviews is also the impression we have had when informally listening to them.

Figure 4. Histogram of the coefficient of variation (*varco*) values among 100 male Czech speakers in reading (dark grey) and in spontaneous speech (light grey).



4. Role of F0 in voice comparison

Fundamental frequency is one of the most frequently used parameters in forensic voice comparison: according to Gold & French (2011), who surveyed 36 forensic phonetic practitioners regarding their analyses, "all respondents routinely measure fundamental frequency" (p. 301), with various indicators of SFF being extracted most commonly. There are definitive positive aspects of F0 analysis: it is not affected by the lexical content of the material so that lexical identity is not required (unlike in the analysis of vowel formants). Although F0 extraction algorithms are certainly not faultless, the median value should not be affected too much by outliers, and especially the baseline value should be most robust in scenarios when manual corrections are not viable.

On the other hand, it is well known that fundamental frequency belongs to the most variable features of a person's voice. Our SFF is affected by a large number of factors, which can be divided into three main groups (Braun, 1995; see also Skarnitzl & Hývlová, 2014 for more details). *Technical factors* include voice disguise; indeed, our voices are very plastic when it comes to the range of F0 we are able to produce. Changing one's SFF has been repeatedly shown to be the most popular disguise strategy in actual casework, and the same has been found in the database of Common Czech (see Růžičková & Skarnitzl, 2017 for more detail). Fortunately, it is generally easy to determine – using auditory analysis – whether a speaker is producing an utterance with his or her natural pitch or whether they are attempting to disguise their voice by shifting the SFF. Braun (1995) mentions the effects of age, smoking or alcohol intoxication as *physiological factors*, while *psychological factors* include especially various affective states or stress. Stress may obviously play a crucial role in forensic settings; in most speakers, stress has been shown to increase SFF (Kirchhübel, Howard & Stedmon, 2011; Giddens, Barron, Byrd-Craven, Clark & Winter, 2013). In addition, Braun (1995) classifies under psychological factors those which we may refer to as *situational*: vocal fatigue, time of the day, level of ambient noise (where increased F0 is caused by the so-called Lombard effect), or speech style.

In light of the many influences on a speaker's F0, it is clear that one must be very cautious when comparing SFF in different voices for forensic purposes. Researchers usually agree that F0 is useful in forensic settings, as long as the physiological and psychological factors of the recordings under investigation are comparable (Hirson et al., 1995; Boss, 1996; Jessen et al., 2005). It is believed that since SFF tends to be higher in unknown recordings, during which speakers usually experience some level of stress, the finding of an SFF value lower than in the known recording obtained during interrogation should speak for non-identity of the speakers (French, 2012, personal communication).

5. Conclusion

The first objective of this paper was to introduce the newly recorded database of Common Czech for forensic purposes, its current level of processing and future plans for its development. The second, and more important objective was to present the first population statistics based on this database; fundamental frequency was chosen as the first parameter.

It is interesting to note once again that, despite all the above-mentioned drawbacks of F0 stemming mostly from the tremendous plasticity of our speech production, F0 still tends to be the most frequently available statistic for a given language. In other words, if there are some population data available for a language, F0 is most probably included in them. Undoubtedly, compiling population statistics of Common Czech for other acoustic parameters typically employed in forensic casework is work which will continue in the future.

ACKNOWLEDGEMENT

This research was supported by the project "Compilation of representative database of Common Czech for forensic purposes" (no. 239515) solved by the second author in 2015 at the Faculty of Arts, Charles University in Prague, financed by Charles University Grant Agency. The first author was also supported by the Charles University project Progres 4, *Language in the shiftings of time, space, and culture*.

REFERENCES

3GPP. (2012). TS 26.071 AMR speech CODEC; General description. Retrieved from http://www.3gpp.org/ftp/specs/archive/26_series/26.071/

Boersma, P. & Weenink, D. (2016). *Praat: Doing Phonetics by Computer (Version 6.0.20)*. Retrieved on September 12, 2016 from www.praat.org

Boss, D. (1996). The problem of F0 and real-life speaker identification: a case study. *Forensic Linguistics*, 3, 155–159.

Braun, A. (1995). Fundamental frequency – How speaker-specific is it? *BEIPHOL 64, Studies in Forensic Phonetics*, 9–23.

Chromý, J. (2014). Demokratizace spisovné češtiny a ideologie jazykové kultury po roce 1948. *Acta Universitatis Carolinae, Philologica*, 3, 71–81. [in Czech]

De Jong, G., McDougall, K. & Nolan, F. (2007). Sound change and speaker identity: An acoustic study. In: Müller, C. (Ed.), *Speaker Classification II*, 130–141. Berlin: Springer.

Hirson, A., French, P. & Howard, D. (1995). Speech fundamental frequency over the telephone and face-to-face: some implications for forensic phonetics. In: Windsor Lewis, J. (Ed.), *Studies in General and English Phonetics: Essays in Honour of Professor J. D. O'Connor*, 230–240. London: Routledge.

Hollien, H. (2002). *Forensic Voice Identification*. San Diego: Academic Press.

Hollien, H. & Hollien, P. A. (1995). Improving aural-perceptual speaker identification techniques. *BEIPHOL 64, Studies in Forensic Phonetics*, 87–97.

Hollien, H., Hollien, P. A. & de Jong, G. (1997). Effects of three parameters on speaking fundamental frequency. *Journal of the Acoustical Society of America*, 102, 2984–2992.

Hudson, T., de Jong, G., McDougall, K., Harrison, P. & Nolan, F. (2007). F0 statistics for 100 young male speakers of Standard Southern British English. *Proceedings of the 16th ICPhS*, 1809–1812.

Hughes, V. & Foulkes, P. (2015). The relevant population in forensic voice comparison: Effects of varying delimitations of social class and age. *Speech Communication*, 66, 218–230.

Jessen, M. (2007). Forensic reference data on articulation rate in German. *Science & Justice*, 47, 50–67.

Jessen, M. (2012). *Phonetische und linguistische Prinzipien des forensischen Stimmenvergleichs*. München: Lincom.

Jessen, M., Köster, O. & Gfroerer, S. (2005). Influence of vocal effort on average and variability of fundamental frequency. *International Journal of Speech, Language and the Law*, 12, 174–213.

Krčmová, M. (2005). Stratifikace současné češtiny. Linguistica Online. Retrieved from http://www.phil.muni.cz/linguistica/art/krcmova/krc-012.pdf [in Czech]

Lindh, J. & Eriksson, A. (2007). Robustness of Long Time Measures of Fundamental Frequency. *Proceedings of Interspeech 2007*, 2025–2028.

McDougall, K., Duckworth, M. & Hudson, T. (2015). Individual and group variation in disfluency features: A cross-accent investigation. *Proceedings of the 18th ICPhS*.

Morrison, G. S. & Ochoa, F. (2012). Database selection for forensic voice comparison. *Proceedings of Odyssey 2012: The Language and Speaker Recognition Workshop*. Singapore: ISCA.

Nolan, F. (1983). *The Phonetic Bases of Speaker Recognition*. Cambridge: Cambridge University Press.

Nolan, F. (1999). Speaker recognition and forensic phonetics. In: Hardcastle, W. J. & Laver, J. (Eds.), *The Handbook of Phonetic Sciences*, 744–767. Oxford: Blackwell Publishers.

Nolan, F., McDougall, K., De Jong, G. & Hudson, T. (2009). The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *International Journal of Speech, Language and the Law*, 16, 31–57.

Pollák, P., Volín, J. & Skarnitzl, R. (2007). HMM-Based Phonetic Segmentation in Praat Environment. *Proceedings of SPECOM 2007*, 537–541. Moscow: MSLU.

R Core Team (2015). *R: A language and environment for statistical computing (version 3.2.2)*. R Foundation for Statistical Computing, Vienna. Retrieved from http://www.R-project.org/.

Reynolds, D. A. & Campbell, W. M. (2008). Text-independent speaker recognition. In: Benesty, J., Sondhi, M. M. & Huang, Y. (Eds.), *Springer Handbook of Speech Processing*, 763–781. Berlin: Springer-Verlag.

Reynolds, D. A., Quatieri, T. F. & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10, 19–41.

Rose, P. (2002). *Forensic Speaker Identification*. London: Taylor & Francis.

Růžičková, A. & Skarnitzl, R. (2017). Voice disguise strategies in Czech male speakers. *AUC Philologica 3, Phonetica Pragensia*, pp. 19–34.

Skarnitzl, R. (2014). Forenzní fonetika. In: Skarnitzl, R. (Ed.), *Fonetická identifikace mluvčího*, pp. 11–20. Praha: Faculty of Arts, Charles University in Prague. [in Czech]

Skarnitzl, R. & Hývlová, D. (2014). Statistický popis hodnot základní frekvence. In: Skarnitzl, R. (Ed.), *Fonetická identifikace mluvčího*, pp. 49–64. Praha: Faculty of Arts, Charles University in Prague. [in Czech]

Skarnitzl, R. & Vaňková, J. (2015). Presenting the population statistics of Common Czech: preliminary F0 results. Presented at IAFPA 2015, Leiden.

Skarnitzl, R. & Volín, J. (2012). Referenční hodnoty vokalických formantů pro mladé dospělé mluvčí standardní češtiny. *Akustické listy*, 18, 7–11. [in Czech]

Vaňková, J. & Bořil, T. (2014). Telefonní přenos. In: Skarnitzl, R. (Ed.), *Fonetická identifikace mluvčího*, pp. 104–115. Praha: Faculty of Arts, Charles University in Prague. [in Czech]

Volín, J. (2007a). *Statistické metody ve fonetickém výzkumu*. Praha: Epocha. [in Czech]

Volín, J. (2007b). Data volume requirements for reliable F0 normalization. In: Vích, R. (Ed.), *17th Czech-German Workshop – Speech Processing*, 62–67. Praha: Czech Academy of Sciences.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag.

RESUMÉ

Identifikace mluvčího je jedním z nejčastějších úloh forenzního fonetika. Při identifikaci mluvčího často porovnáváme nahrávky hlasů dvou nebo více mluvčích a cílem srovnání je posoudit jejich podobnost, ale také typičnost. Je zřejmé, že posuzování typičnosti hlasů a rozdílů mezi nimi je třeba mít k dispozici informace o populaci mluvčích. Pořizování databází a kompilace statistik pro použití ve forenzně fonetické praxi proto patří k důležitým aspektům současného výzkumu. Článek popisuje databázi obecné češtiny, která byla pořízena právě s cílem, aby sloužila jako referenční databáze pro forenzní účely, a zároveň představuje první sadu populačních statistik týkajících se základní frekvence (F0). Prezentované údaje jsou založené na čtené i spontánní řeči. Výsledky potvrzují, že tzv. základní hladina F0 navržená Lindhem a Erikssonem (2007) vystihuje idiosynkratičnost F0 lépe než aritmetický průměr či medián. Naznačují také, že velká část mluvčích ve spontánním projevu hovoří monotónněji, tedy vykazuje nižší intonační rozpětí. Článek také rozebírá roli základní frekvence ve srovnávání mluvčího.

Radek Skarnitzl, Jitka Vaňková
Institute of Phonetics
Faculty of Arts, Charles University
radek.skarnitzl@ff.cuni.cz

VOICE DISGUISE STRATEGIES IN CZECH MALE SPEAKERS

ALŽBĚTA RŮŽIČKOVÁ AND RADEK SKARNITZL

ABSTRACT

Voice comparison in forensic phonetic casework requires the assessment of the similarity of the target voices. This task may be impeded by the speakers' attempt to disguise their voice. The objective of this study is to map the strategies of voice disguise as employed by 100 native speakers of Common Czech. In agreement with findings in other languages, changes in speaking fundamental frequency appeared in most speakers, and phonatory modifications were the second most frequent type of disguise. Other strategies included modifications of the resonance characteristics of the voice, melodic and temporal changes. Recognition difficulty and naturalness of the disguise were also assessed. In addition, several acoustic parameters in the natural and disguised condition were compared in 15 speakers whose disguise rendered recognition more difficult.

Key words: voice, voice disguise, forensic phonetics, speaker identification, Czech

1. Introduction

The most typical task of a forensic phonetician concerns the recording of a perpetrator of a criminal act – often obtained from mobile telephone interception – which has to be compared with a recording of a detained suspect. In this procedure, which is called voice comparison, the aim is to arrive at a statement of probability with which the two recordings originate from the same speaker. If the perpetrator is aware of the option of their communication being monitored, he or she may attempt to disguise their voice, which can render voice comparison difficult or even impossible.

Current statistics of the occurrence of voice disguise are, unfortunately, not available. Data from Germany indicate that, overall, such cases are generally not very frequent: in the years 1988–1995, attempts at voice disguise appeared in less than 5 per cent of all cases investigated at Trier University (Masthoff, 1996). However, the frequency of occurrence was reported as considerably higher in cases of abductions, extortions, sexual harassment, or hoax calls (Künzel, 2000), with disguise attempts identified as frequently as

https://doi.org/10.14712/24646830.2017.30

in 69% of blackmail cases (Masthoff, 1996). It may be expected that, with the increase in mobile phone communication, the proportion of voice disguise will continue to increase. Indeed, in a more recent study, Braun (2006) examined 175 forensic cases and reported the occurrence of disguise in 22.9% of them, with the number highest in harassment and extortion cases. The possibility of voice disguise is therefore something which must be kept in mind at all times when comparing voices in forensic phonetic practice.

This paper is interested in the strategies which speakers employ when attempting to disguise their voice. One type of voice disguise which will not be addressed here in more detail is electronic disguise, in which the speaker makes use of a specialized device; this type of disguise used to be quite infrequent (less than 1% of all cases according to Masthoff, 1996) but seems to be on the rise (Clark & Foulkes, 2007; see also Perrot et al., 2007). The majority of voice disguise cases thus still involves speakers exploiting the natural potential for variability in their speech production (Skarnitzl, 2016) and changing some characteristics of their speech.

Before discussing the individual strategies which speakers use to disguise their voice, let us point out the results of Masthoff's study, in which speakers were instructed to disguise their voice in any way they could think of. Masthoff shows that speakers typically chose to change only one or two parameters in their speech. A single parameter was altered by 55% of the subjects. This corresponds to the experience of forensic phoneticians in the Czech Republic (Svobodová & Voříšek, 2014), who also report relatively primitive voice disguise attempts. As Masthoff (1996) states, the fact that the speakers do not use more complex strategies may be caused by the speaker's need to formulate utterances simultaneously with attempting voice disguise, which requires a relatively high cognitive effort.

In the following sections, the most frequent disguise strategies reported in literature will be described; where available, we will also discuss the effect of the particular strategies on recognition of the speaker's voice by listeners.

1.1 Fundamental frequency changes

The most frequent voice disguise strategy has been repeatedly shown to be changes of fundamental frequency (F0). When discussing the overall level of a speaker's fundamental frequency, we talk about speaking fundamental frequency (SFF). Modifications of SFF are advantageous for the speaker because they are easy to perform, they do not cause incomprehensibility of the utterance (and therefore do not restrain the transfer of information), and they are sufficiently effective for disguising the speaker's identity (Künzel, 2000).

According to Künzel's (2000) experiment, speakers with a higher natural SFF tend to raise it even more, while speakers with a lower natural SFF tend to lower it, sometimes shifting to creaky voice. In his study, in which speakers were asked to speak with a higher and lower SFF, females made less significant changes whereas males in general modified SFF more prominently and they shifted to falsetto more frequently (see section 1.2 for more information on creaky voice and falsetto).

Extreme cases of F0 modification also lead to changes of vowel formants; although F0 and formants are often said to be independent of each other, extreme F0 values are associated with changes in the vertical position of the larynx. A raised larynx due to high F0 results in a shorter vocal tract and hence higher resonance frequencies (formants), and vice versa.

Künzel (2000) also notes that F0 modifications lead to the decrease of articulation rate and to the increase in the occurrence of pauses and their duration, probably as a result of the speaker's concentration on maintaining the disguise strategy.

The effect of SFF changes on the listeners' ability to identify speakers was studied by one of the first researchers in this field of study, the psychologist Frances McGehee (1937; cited in Eriksson, 2010). She experimentally tested the accuracy of voice line-ups when F0 had been modified and found that recognition dropped from 80% (when listeners heard the speakers' natural voice) to 63% (when they first heard the speaker's disguised voice). Other studies focused on voice identification when the change in SFF was a by-product of a different type of phonation; these will be discussed in the following section.

1.2 Phonatory modifications

The most frequent modifications to modal phonation which have been reported in forensic casework, creaky voice and falsetto, also include a change in F0. Creaky voice is a result of a complex laryngeal adjustment which involves tighter adduction of the arytenoid cartilages and low-frequency and low-amplitude vocal fold vibration only in the front portion of the glottis. The effectiveness of creaky voice disguise was examined by Hirson and Duckworth (1995) who compared the performance of expert and lay listeners. They showed that while speaker identification was high when the speakers used modal phonation, 99% for phoneticians and 93% for non-phoneticians, performance dropped significantly when the speakers used creaky voice: to 73% for phoneticians and to 51% for non-phoneticians. The results therefore suggest that phonetically trained listeners perform better in voice line-ups, even when the speakers disguise their voice.

The second phonatory modification, falsetto, involves a non-modal pattern of vocal fold vibration which is characterized by high longitudinal tension and high-frequency vibration only in the front portion of the glottis. The effectiveness of falsetto as a voice disguise strategy was studied by Wagner and Köster (1999), who presented recordings of familiar and unfamiliar speakers to listeners; the speakers used both modal phonation and falsetto. The results show how detrimental falsetto was to recognition: while the listeners were able to identify 97% of the familiar speakers in the undisguised condition, only 4% of the same speakers were correctly recognized when they used falsetto. The authors highlight the importance of using phonetically trained experts for forensic phonetic analyses.

Another type of non-modal phonation, called pressed or tense voice, involves a high compression of the vocal folds in the medial portions of the vocal folds; this voice is sometimes accompanied by the vibration of the false vocal folds, by the so-called ventricular phonation. To our knowledge, pressed voice has not been examined from the perspective of its effectiveness as a voice disguise strategy; its use in forensic settings has, however, been reported (Künzel, 2000).

Finally, speakers disguising their voice may decide to eliminate phonation entirely and whisper. It is to be expected that whispered speech lacks some crucial speaker-specific information, and speaker identification will thus be compromised. The effect of whisper on voice line-ups was investigated by Orchard and Yarmey (1995) who found that line-ups were most successful when the listeners had heard the speaker's natural speech in both speech samples, less accurate when the listeners had heard the speaker's whisper on both samples, and least accurate when the listeners first heard the speaker's whisper and then were asked to identify the speaker based on his normal speech. Speaker identification in normal and whispered speech has also been examined by Bartle and Dellwo (2015) who, similarly, report a drop in correct identifications in whispered speech, and more so for naïve listeners than for forensic phoneticians.

1.3 Resonance modifications

Apart from voice disguise strategies which exploit changes in the frequency or manner of vocal fold vibration, speakers can modify the resonance characteristics of the speech signal in several ways. First, we can mention changes to the supralaryngeal long-term settings (Laver, 1980; Skarnitzl, 2016). These include palatalization, labialization or pharyngealization (in other words, quasi-permanent shifts away from the neutral position of the tongue), as well as hyper- and hyponasality (habitual opening of the velopharyngeal port and speaking "through the nose" in the former case, and pinching one's nose so as to prevent nasal airflow in the latter). To the best of our knowledge, the effectiveness of adopting these long-term setting modifications for voice disguise purposes has not been studied.

The second type of resonance changes involves the speaker inserting a foreign object into his or her vocal tract or holding it in front of their mouth (Künzel, 2000). Figueiredo and Britto (1996) investigated the effect on the voice of holding a pen between the front teeth: the articulation of speech sounds is restricted due to the fixed position of the jaw, spreading of the lips and retraction of the tongue. Various speech segments are, understandably, affected in different degrees; it is therefore not possible to make a general prediction as to the acoustic effect of a foreign object in the oral cavity. The authors point out that any disguise strategy which modifies the speaker's supralaryngeal characteristics impedes identification more than phonatory modifications, because the latter changes preserve most dialectal and segmental features in speech. The reason for the low predictability of supralaryngeal changes consists mainly in the complex interaction between the restriction caused by the object in the speaker's mouth on the one hand and the speaker's natural articulation manners on the other hand (Figueiredo & Britto, 1996). In more primitive ways of voice disguise, whose impact on the resonance characteristics is relatively low, speakers often use a handkerchief placed between their mouth and the microphone (Svobodová & Voříšek, 2014).

1.4 Dialect or foreign accent imitation

A dialect can be so strong a component of identity that it can obscure, or override other features of the speaker's voice (Eriksson, 2010). It is therefore not surprising that imitating another dialect belongs to frequently encountered strategies in forensic phonetic casework. The key question is, naturally, the authenticity of the imitation. Markham (1999) studied eight speakers of Southern Swedish, each imitating three regional accents of Swedish, and found that some of the speakers were able to achieve a consistently authentic impression and conceal their own identity.

Another frequent strategy of changing one's pronunciation concerns the imitation of a foreign accent (Masthoff, 1996), even in the Czech environment (Svobodová & Voříšek, 2014). It is obvious that as with dialect imitation, authenticity must be maintained for disguise to be successful, and speakers are not always able to achieve that (Neuhauser, 2008). Neuhauser and Simpson (2007) investigated the ability of native German listeners to identify authentic and imitated French and American English accents in German. Their results show, surprisingly, that German listeners were better able to identify imitated accents by German speakers than authentic non-native accents (uttered by native speakers of the respective languages); the authors hypothesize that the German imitators made use of stereotypical pronunciation patterns which were, in turn, picked up by the listeners.

1.5 Voice disguise strategies and this study

When we compare the overview studies which have been quoted in the previous sections (Masthoff, 1996; Künzel, 2000; Braun, 2006), the emerging picture of voice disguise strategies tends to be quite similar. The majority of speakers who attempt at disguise modify one, maximum two parameters in their speech. Voice disguise mostly involves a change in the phonatory behaviour, i.e., in some characteristics of the voice: according to Braun (2006), this concerned 65% of all disguise attempts. Most frequently, speaking fundamental frequency is raised (possibly switching into falsetto), sometimes also lowered (possibly switching into creaky phonation); the former seems to be preferred by male speakers, the latter by females. Pressed voice and whisper are also relatively frequent. Other strategies include denasalization (achieved by pinching one's nose), the imitation of a foreign accent, or speaking with a handkerchief in front of the mouth. The same strategies have been witnessed in the Czech forensic environment (Svobodová & Voříšek, 2014).

Investigations of voice disguise also clearly indicate that speakers differ in their ability to effectively and consistently (without "gaps" in the disguise) conceal their identity (Masthoff, 1996; Neuhauser, 2008; Vyhnálková, 2013), just as listeners differ in their ability to identify speakers who manipulate their voices (Vyhnálková, 2013).

This paper follows up on Vyhnálková's pilot study and examines voice disguise strategies in a large corpus of 100 male speakers of Common Czech (Krčmová, 2005; Chromý, 2014), a nonstandard but supraregional dialect of the Czech language used in everyday communication by many speakers not only in informal but even in slightly formal situations.

2. Auditory mapping of voice disguise strategies

2.1 Material

The mapping of voice disguise strategies was conducted on the Database of Common Czech, which contains recordings of 100 male speakers aged between 19 and 50 (mean age: 25.6 years) and represents a reference database used for forensic purposes; the database only features male speakers precisely because most forensic material is produced by males. The recordings were acquired in a quiet environment, using a portable recorder Edirol R09, with 48-kHz sampling frequency. The speakers performed several speaking tasks which involved several speech styles (see Skarnitzl & Vaňková, 2017 for more information on the corpus); this study is based on the analysis of two reading tasks.

In the first task, the subjects were asked to read in their natural voice a phonetically rich text of 150 words, lasting approximately one minute. For the second, voice disguise task, the speakers were instructed to imagine that they were criminals who had to report to their boss, while suspecting the call might be monitored. They were therefore told to change their voice so they would not be identifiable based on the recording. They were given sufficient time to devise a strategy to disguise their voice. The two texts differed but contained some identical phrases.

2.2 Procedure

The first objective was to perform auditory analyses of both types of recordings – natural and disguised – in order to map the range of voice disguise strategies used by the speakers; this step was performed by the first author. The careful, repeated auditory comparison of the natural and disguised recordings yielded a list of disguise strategies corresponding to the following categories:

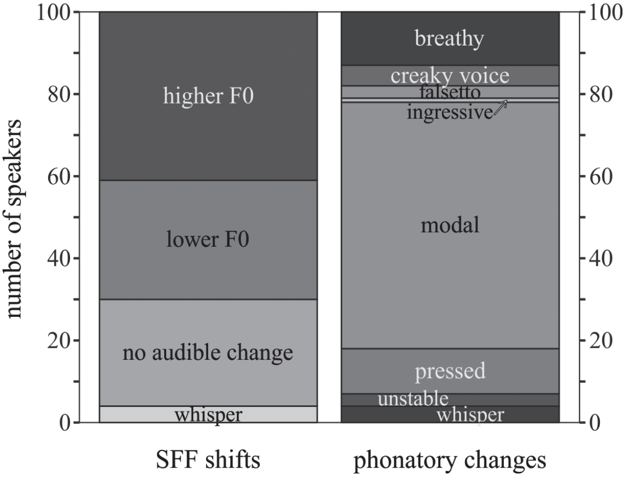
- speaking fundamental frequency (pitch level) modifications (higher, lower, no audible change)
- phonatory modifications (modal, creaky, pressed, breathy phonation, whisper)
- speech rate (faster, slower, no audible change)
- any other notable strategies (nasalization and other resonance changes, etc.)

Apart from noting the disguise strategy, we subjectively evaluated another two criteria: the difficulty of recognition of the speaker based on his natural and disguised voice (easy, medium, difficult), and the naturalness of the speaker's disguised voice (natural, unnatural). The second criterion was thus concerned whether it was easily noticeable that the speaker was disguising their voice or not. For instance, changes in articulation sounded natural in most cases, and a hearer might therefore not identify such speech as intentionally modified, whereas prominent SFF changes would be regarded as an obvious disguise attempt. The natural and disguised production were compared within the same speaker, because the objective of this step was to separate speakers who chose rather simple and ineffective ways of disguise from those who used more complex strategies and whose voice disguise strategies were suitable for a more thorough analysis.

2.3 Results

As expected, based on the literature reviewed above, the majority of speakers tended to shift their speaking fundamental frequency (see the left panel of Figure 1). In most cases, SFF was audibly higher (41 of the 100 speakers), with 3 of them even shifting to falsetto. 29 speakers lowered their SFF, and 5 of them shifted into creaky voice. 26 speakers' pitch level remained without any audible change. The remaining 4 speakers disguised their voice using whisper, so that F0 was eliminated altogether. Of the speakers who used SFF shift as voice disguise, 17 only changed this parameter (6 speakers raised and 11 lowered their SFF).

Figure 1. Disguise strategies in the form of speaking fundamental frequency (SFF) shifts (on the left) and phonatory modifications (right).



As can be seen in the right panel of Figure 1, the manner of phonation was modified only by a minority of the speakers: 58 of them adhered to their modal phonation while disguising their voice. Of the remaining 42 speakers, 13 employed breathy phonation and 13 pressed phonation, 5 spoke in a creaky voice, 3 shifted into falsetto, and 4 speakers used whisper. 1 speaker employed ingressive phonation. In 3 speakers, the manner of phonation was unstable: they used more types within their disguised recording. Interestingly, a phonatory modification was the single changed parameter only in 4 speakers: 2 of them employed pressed phonation, 1 speaker used ingressive phonation, and 1 speaker's phonation was unstable (falsetto, creaky voice, as well as modal phonation appeared in his disguised speech). Some of the speakers' disguise strategies involved more phonatory modifications, for example a combination of pressed and creaky voice; these more sophisticated techniques will be mentioned separately in section 4.

Speech rate was audibly modified even less frequently than phonation: it remained unchanged in the disguised condition of 86 of the 100 speakers. Lower speech rate appeared in 11 speakers, only 3 speakers decided to speak faster. None of the speakers employed a change in speech rate as the only modification.

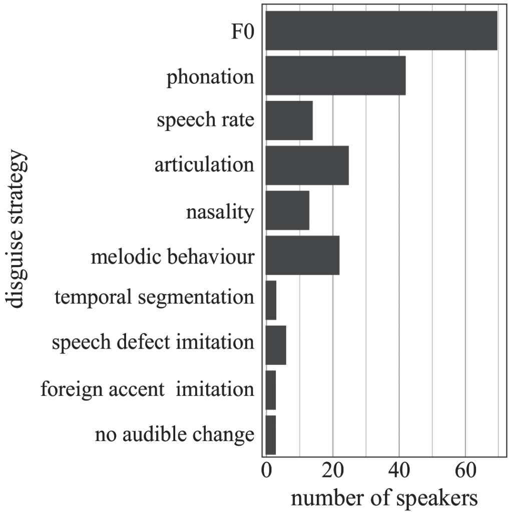
39 speakers used some kind of modification of the resonance characteristics of their speech. Quite popular among these were changes in nasality: 11 speakers used hypernasality as a long-term setting, while 2 speakers pinched their nose and sounded de-nasalized. 4 speakers modified their long-term lingual setting to pharyngealization. In 10 cases, we observed changes in the articulation of vowels: anteriorization in 3 cases, posteriorization in 2 cases, higher articulation in 3 cases and lower articulation in 2 cases.

6 speakers imitated one or more speech defects, and 3 speakers imitated a foreign accent (dialect imitation did not occur in our sample, though).

25 speakers modified their prosodic behaviour: 13 of these changed the realization of conclusive falling tones, while 9 spoke in a monotonous voice. Another 3 speakers used a lot of pauses and hesitation markers in their speech.

Finally, it remains to be pointed out that only 3 of the 100 speakers failed to perform any audible changes in their voice disguise. The strategies selected by the speakers in our study are summarized in Figure 2; note that as some speakers disguised their voice in more ways at the same time, the total number of disguise strategies exceeds the number of speakers.

Figure 2. Summary of disguise strategies, as identified by means of careful listening.



Next, we will turn to the two additional criteria that we rated, recognition difficulty and naturalness (see above in section 2.2). The ratings are summarized in Table 1. In terms of the difficulty of recognition, it was usually obvious that the given speaker's disguised voice was produced by the same individual as his undisguised voice: 53 speakers' disguise was thus rated as easy to recognize. In 30 of these speakers, the disguise did not sound natural, while the other 23 ones achieved naturally sounding disguised speech.

For 32 speakers, we rated the difficulty of recognition as medium; the match between their disguised and undisguised voices was not entirely straightforward at first "sight", but careful listening would probably lead to a successful recognition without any serious troubles. Within this group of speakers, 22 of them sounded unnatural in their disguised condition, and only 10 speakers' disguised voices were rated as sounding natural.

High difficulty of matching the disguised speech to the natural speech of the same speaker was assigned in 15 cases; 7 of the speakers' disguise sounded unnatural and 8 natural.

As is summarized in the last column of Table 1, 59 speakers' disguised voices sounded intentionally modified (i.e., the speakers did not sound natural), while 41 speakers' disguise gave the impression of a more or less natural (not intentionally modified) speech.

Table 1. Summary of the naturalness and recognition difficulty of disguise strategies, as identified by means of careful listening.

	recognition difficulty			
	low	medium	high	total
natural disguise	23	10	8	41
unnatural disguise	30	22	7	59
total	53	32	15	100

3. Acoustic analyses

3.1 Method

Based on the outcomes of the auditory analyses presented in section 2.3, we selected 15 speakers whose recognition was judged as difficult. Subsequently, these 15 speakers' natural and disguised voices were analyzed acoustically.

Segment boundaries of the 30 target recordings (natural and disguised for 15 speakers) were aligned automatically using Prague Labeller (Pollák et al., 2007) and corrected manually. Acoustic analyses were then performed on the entire recordings, as well as on individual sounds, as described below.

As for speaking fundamental frequency, we extracted raw F0 values automatically every 10 msec using autocorrelation in Praat (Boersma & Weenink, 2015); as some of the speakers used falsetto or a notably high SFF, the extraction range was set to 60–450 Hz. Speaking fundamental frequency was then quantified in several ways (*cf.* Skarnitzl & Vaňková, 2017): we calculated the mean and median value of F0, as well as the more robust F0 baseline (Lindh & Eriksson, 2007; see also Skarnitzl & Hývlová, 2014).

The following analyses regarding phonatory modifications were conducted on sounds resampled to 16 kHz. First of all, harmonicity (harmonics-to-noise ratio, HNR), which is related to the degree of noise in the spectrum (Yumoto, Gould & Baer, 1982), was extracted for each vowel using the default settings in Praat (Boersma, 1993). It was necessary to deal with those tokens where Praat yielded an "undefined" value; it did not seem advantageous to exclude such tokens since the absence of periodicity reflected the disguise strategy employed by some speakers. With HNR, the lowest detected value was –3.6 dB; undefined values were thus replaced with –10 dB (*cf.* Skarnitzl, 2011: 223). The other two parameters expressing some aspects of voice quality were jitter and shimmer as measures of voicing irregularity; these were also extracted from Praat, using the default values.

The extent of (at least some) articulatory modifications was assessed by means of comparing long term formant distributions (LTFs; see Nolan & Grigoras, 2005). Formant values (F1–F3) were extracted in the 0–5 kHz range every 10 msec from vowel and non-nasal sonorant segments, using the Burg algorithm implemented in Praat.

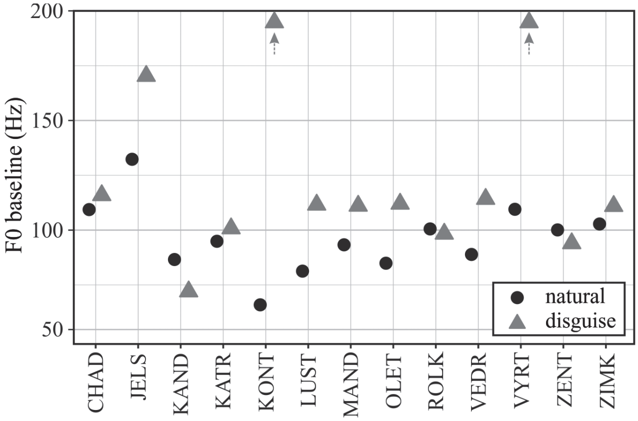
The last parameter we wanted to objectify was speech rate. We therefore measured for each speaker's masked natural and disguised reading, in syllables per second, speech rate and articulation rate (the latter being speech rate stripped of pauses, hesitations etc.).

Statistical analyses were performed in R (R Core Team, 2015) using the package *effects* (Fox, 2003) and visualized using the package *ggplot2* (Wickham, 2009).

3.2 Results

It is obvious from the results of the auditory assessment presented in section 2.3 that speaking fundamental frequency (SFF) belonged to the most frequently modified parameters. Figure 3 confirms the perceptual data: in many of the 15 speakers we can see a clear change in SFF, although SFF is expressed as F0 baseline, which is relatively more robust to behavioural changes. It is clear, however, that extreme pitch shifts will translate into baseline shifts as well. The differences in SFF between the natural and disguised condition would have been considerably more pronounced if SFF were depicted using the median value. Note that the comparison for speakers KUCK and MAKN is missing, as they employed whisper to disguise their voice, and any F0 values detected in the disguise condition were spurious.

Figure 3. Comparison of speaking fundamental frequency (expressed as the baseline) in natural and disguised speech.



Next, let us turn to the parameters pertaining to voice quality. Figure 4 shows a comparison of mean harmonicity in the natural and disguise condition for each speaker. Harmonicity seems to be sensitive at least to some audible changes in phonation: the speakers who changed their voice are indicated in italics in the figure, and greater differences in HNR can be observed in most of these speakers. It is worth commenting on speaker KATR, whose mean harmonicity is considerably higher in the disguise mode but was not marked as having modified his voice quality. What he did change was the resonance characteristics of his voice by means of pharyngealization and lowering his larynx, with the result of his voice sounding more resonant; hence higher harmonicity. Some of the other speakers whose disguised speech yielded higher HNR values, KONT and VYRT, spoke in falsetto (*cf.* Fig. 3), which also makes the speech wave more regular and thus represents a relative increase in tonal components. On the other hand, the lowest HNR values in the disguise condition (KUCK, MAKN, ROLK, ZENT) are the result of breathy voice, whisper, or a voice which combines creaky and breathy quality.

Figure 4. Comparison of mean harmonicity in natural and disguised speech; speakers who were identified as having modified their phonation behaviour are indicated in italics.

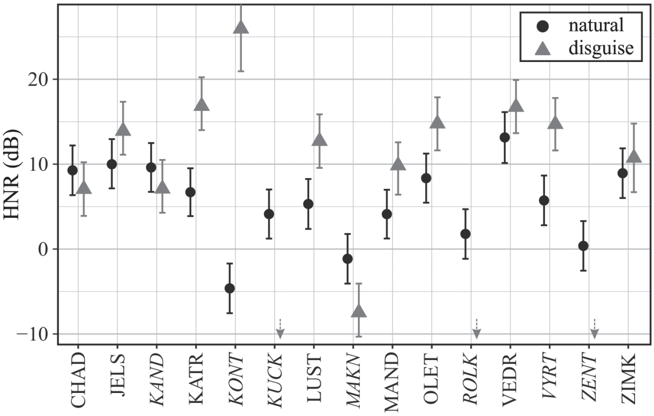
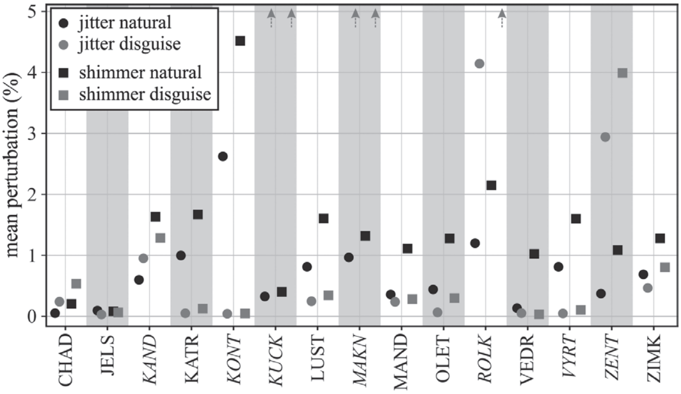


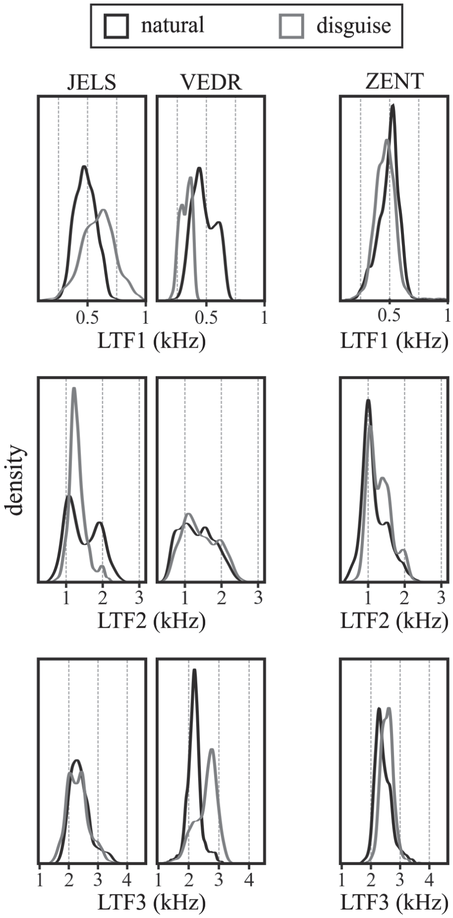
Figure 5 shows results for the F0 perturbation measures, jitter and shimmer. It must be kept in mind that the values are quite low since they were extracted from vowels in natural speech, which are quite short. As with HNR above, the speakers indicated in italics have audibly modified their voice quality. Again, in most of these speakers, the disguise conditions (in gray) are associated with considerably higher values than natural conditions (in black).

Figure 5. Comparison of mean jitter and shimmer in natural and disguised speech; speakers who were identified as having modified their phonation behaviour are indicated in italics. The arrows at the top of some bars mean a higher mean value due to the replacement of undefined values (see section 3.1).



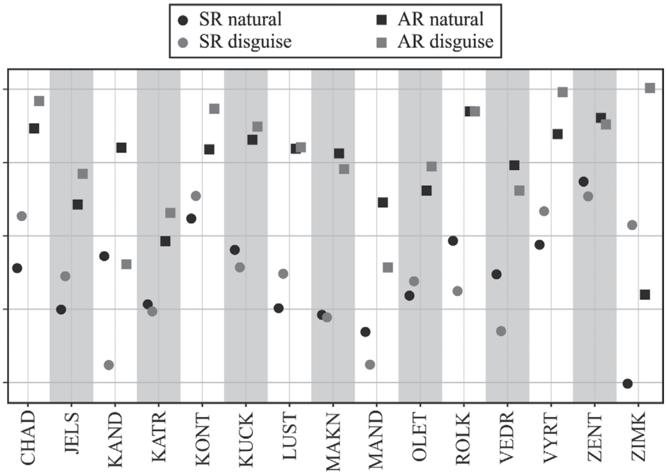
Let us now move from characteristics pertaining to the voice source to changes in supraglottal resonances. The auditory mapping presented in section 2.3 revealed that several speakers did make changes to their articulation. We therefore compared long-term formant distributions in these speakers whose articulatory characteristics sounded different in the disguise condition with LTFs of those where such a difference was not identified. Partial results of this comparison are shown in Figure 6, with distributions of F1–F3 for two speakers from the former and one speaker from the latter group. It can be seen that, indeed, the formant distributions in natural and disguised speech are much more similar – especially in terms of the location of the primary peak – in the speaker on the right, where no articulatory modifications were detected, than in the two speakers on the left, whose disguise strategy included articulatory shifts.

Figure 6. Comparison of LTF1–3 in natural and disguised speech for two speakers whose articulation was audibly different (on the left) and for one speaker whose articulation did not sound changed (on the right).



The final analysis concerns speech rate (SR) and articulation rate (AR). In the impressionistic assessment of speech rate changes in section 2.3, no salient changes have been noted. Figure 7 confirms that changes exceeding 1 syllable per second are quite rare (speakers KAND and ZIMK in both SR and AR, with the former speaking slower in the disguise condition while the latter speaking faster).

Figure 7. Speech rate (SR) and articulation rate (AR) in natural and disguised speech, in syll/s.



4. Interesting cases of voice disguise

As was mentioned in section 3.2, there were 3 out of the 100 speakers who did not perform any modifications under the disguise condition. Another 31 speakers performed only one change to their speech. In this section, we are going to survey several speakers from the other end of the scale, who demonstrated a greater degree of imagination when choosing the strategies for voice disguise. All of them were quite difficult to recognize when compared with their natural speech production. Sound examples are provided on the accompanying webpage, http://fonetika.ff.cuni.cz/vyzkum/materialy/maskovani/.

First, let us mention speaker ROLK, who combined a strongly breathy voice quality (*cf.* Fig. 4) with imitating a foreign accent; we can assume that he was aiming at a Russian accent. This was manifested by a closer (higher) articulation of vowels, by quite subtle shifts of consonantal articulation (especially the postalveolar fricatives and the fricative trill ř), and by changes in word stress and in rhythmical patterning in general. Along with high consistency, this all led to a naturally sounding disguise which would render speaker identification rather difficult.

Another successful imitation strategy was employed by speaker ZIMK, who decided to imitate the current Czech president Zeman. The strategies included melodic and rhythmic changes, with individual words or short groups of words being chopped off, sometimes even using preglottalization (e.g., [ʔʒɛ ˈʔr̝iːɦovo͡u ˈʔsɛm]; see Skarnitzl & Machač, 2010), as well as segmental changes such as shortening of vowels or more close articulation.

Speaker OLET managed to disguise his voice very effectively by modifying a single parameter: the vertical position of the larynx. His laryngeal lowering still sounds relatively natural, but its degree is suggested by changes in vowel formants: for instance, in the word *vlastně*, the F1 in [ɛ] dropped by 0.67 ERB and the F2 by over 2 ERB.

Modifications of articulation were the dominant disguise strategy of speaker VEDR who, apart from raising SFF, imitated several speech defects. The lateral alveolar approximant [l] was pronounced with no alveolar contact, only with slight raising of the tongue tip, and therefore vocalized. The alveolar trill [r] was in most cases realized in a very similar way, but sometimes also as a uvular fricative. The fricative trill was usually pronounced as a fricative, without the initial trill. Apart from these articulatory changes, the speaker also raised his SFF and spoke with very prominent lengthening of phrase-final syllables.

The last speaker to be mentioned here is LUST, who also exploited a range of changes to disguise his voice. Apart from a slightly higher SFF (*cf.* Fig. 3), he also imitated a speech defect, specifically more anterior (dentalized) articulation of alveolar fricatives and a shift in the Czech fricative trill ř. On top of that, he also employed melodic and rhythmic changes.

5. Discussion

The aim of this study was to examine the strategies of voice disguise employed by 100 native speakers of Czech. Not surprisingly, changes in speaking fundamental frequency appeared in 70% of the speakers. Phonatory modifications as a group were the second most frequent type of disguise, with creaky and pressed voice each appearing in 13 speakers. For each speaker, we also noted the naturalness of his speech in the disguise mode, as well as how difficult it was to recognize the speaker in his disguise. Apart from mapping disguise strategies in the entire database, we performed acoustic analyses in 15 speakers who managed to change their voice in such a way that recognition would be difficult (section 3) and introduced the most interesting disguise strategies of six selected speakers.

Based on the results of this study, it does not seem that we could easily identify one strategy or a combination of more strategies which are most effective when one attempts to disguise their voice. In section 2.3, we mentioned that there were 8 speakers whose disguise sounded natural and was difficult to assign to their natural speech. No clear pattern emerges even if we examine only these speakers: 6 of them did change their SFF but 2 did not; 7 of them performed some kind of articulatory modification (nasalization, closed jaw setting, speech defects); phonation was changed in only 2 of these speakers.

Voice disguise strategies represent a very interesting research topic, not only due to its practical implications in forensic phonetic casework, but also due to the way in which it illustrates the astounding plasticity of the human speech production mechanism. It is clear that investigations of voice disguise will continue in the future, as will further surveys of the Common Czech database for other forensically relevant properties of speech.

ACKNOWLEDGEMENTS

The main author was supported by the Internal Grants of the Faculty of Arts programme, registration number FF/VG/2017/83. The second author was supported by the Charles University project Progres 4, *Language in the shiftings of time, space, and culture*.

REFERENCES

Bartle, A. & Dellwo, V. (2015). Auditory speaker discrimination by forensic phoneticians and naive listeners in voiced and whispered speech. *The International Journal of Speech, Language and the Law*, 22, pp. 229–248.

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *IFA Proceedings*, 17, pp. 97–110.

Boersma, P. & Weenink, D. (2015). *Praat: doing phonetics by computer (Version 6.0)*. Retrieved from http://www.praat.org

Braun, A. (2006). Stimmverstellung und Stimmenimitation in der forensischen Sprechererkennung. In: Kopfermann, T. (Ed.), *Das Phänomen Stimme: Imitation und Identität*, pp. 177–181. Röhrig: St. Ingbert.

Chromý, J. (2014). Demokratizace spisovné češtiny a ideologie jazykové kultury po roce 1948 [The democratization of Standard Czech and the ideology of language culture after 1948]. *Acta Universitatis Carolinae – Philologica*, 3, pp. 71–81.

Clark, J. & Foulkes, P. (2007). Identification of voices in electronically disguised speech. *The International Journal of Speech, Language and the Law*, 14, pp. 195–221.

Eriksson, A. (2010). The Disguised Voice: Imitating Accents or Speech Styles and Impersonating Individuals. In: Llamas, C. & Watt, D. (Eds.), *Language and Identities*, pp. 86–96. Edinburgh: Edinburgh University Press.

Figueiredo, R. M. & Britto, H. S. (1996). A report on the acoustic effects of one type of disguise. *Forensic Linguistics*, 3, pp. 168–175.

Fox, J. (2003). Effect displays in R for generalised linear models. *Journal of Statistical Software*, 8(15), pp. 1–27.

Hirson, A. & Duckworth, M. (1995). Forensic implications of vocal creak as voice disguise. In: *BEIPHOL 64*, *Studies in Forensic Phonetics*, pp. 67–76.

Krčmová, M. (2005). Stratifikace současné češtiny [Stratification of contemporary Czech]. *Linguistica Online*. Retrieved from http://www.phil.muni.cz/linguistica/art/krcmova/krc-012.pdf

Künzel, H. J. (2000). Effects of voice disguise on speaking fundamental frequency. *Forensic Linguistics*, 7, pp. 149–179.

Laver, J. (1980). *The Phonetic Description of Voice Quality*. Cambridge: Cambridge University Press.

Lindh, J. & Eriksson, A. (2007). Robustness of long time measures of fundamental frequency. In: *Proceedings of Interspeech 2007*, pp. 2025–2028.

Masthoff, H. (1996). A report on a voice disguise experiment. *Forensic Linguistics*, 3, pp. 160–167.

Neuhauser, S. (2008). Voice disguise using a foreign accent: phonetic and linguistic variation. *The International Journal of Speech, Language and the Law*, 15, pp. 131–159.

Neuhauser, S. & Simpson, A. P. (2007). Imitated or authentic? Listeners' judgements of foreign accents. In: *Proceedings of 16th ICPhS*, pp. 1805–1808.

Nolan, F. & Grigoras, C. (2005). A case for formant analysis in forensic speaker identification. *International Journal of Speech, Language and the Law*, 12, pp. 143–173.

Nolan, F., McDougall, K., De Jong, G. & Hudson, T. (2009). The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *International Journal of Speech, Language and the Law*, 16, pp. 31–57.

Orchard, T. L. & Yarmey, A. D. (1995). The Effects of Whispers, Voice-Sample Duration, and Voice Distinctiveness on Criminal Speaker Identification. *Applied Cognitive Psychology*, 9, pp. 249–260.

Perrot, P., Aversano, G. & Chollet, G. (2007). Voice disguise and automatic detection: review and perspectives. In: Stylianou, Y., Faundez-Zanny, M. & Esposito, A. (Eds.), *Workshop on Nonlinear Speech Processing 2005, LNCS 4391*, pp. 101–117. Berlin: Springer Verlag.

Pollák, P., Volín, J. & Skarnitzl, R. (2007). HMM-based phonetic segmentation in Praat environment. In: *Proceedings of the XIIth International Conference "Speech and computer – SPECOM 2007"*, pp. 537–541.

R Core Team (2015). *R: A language and environment for statistical computing (version 3.2.2)*. R Foundation for Statistical Computing, Vienna. Retrieved from http://www.R-project.org

Skarnitzl, R. (2011). *Znělostní kontrast nejen v češtině* [*Voicing Contrast Not Only in Czech*]. Praha: Epocha.

Skarnitzl, R. (2016). Co dokáže náš hlas? Fonetický pohled na variabilitu řečové produkce [What is our voice capable of? Phonetic perspective on the variability of speech production]. *Slovo a smysl*, 26, pp. 95–113.

Skarnitzl, R. & Hývlová, D. (2014). Statistický popis hodnot základní frekvence [Statistical description of fundamental frequency values]. In: Skarnitzl, R. (Ed.), *Fonetická identifikace mluvčího* [*Phonetic Speaker Identification*], pp. 49–64. Praha: Faculty of Arts, Charles University in Prague.

Skarnitzl, R. & Machač, P. (2010). Domain-initial coordination of phonation and articulation in Czech radio speech. *Acta Universitatis Carolinae – Philologica*, 1/2009, pp. 21–35.

Skarnitzl, R. & Vaňková, J. (2017). Fundamental frequency statistics for male speakers of Common Czech. *Acta Universitatis Carolinae – Philologica*, 3, pp. 7–17.

Svobodová, M. & Voříšek, L. (2014). Identifikace mluvčích z pohledu autentické kriminalistické praxe v České republice [Speaker identification from the perspective authentic criminological practice in the Czech Republic]. In: Skarnitzl, R. (Ed.), *Fonetická identifikace mluvčího* [*Phonetic Speaker Identification*], pp. 136–144. Praha: Faculty of Arts, Charles University.

Vyhnálková, L. (2013). *Vliv vzdělání na schopnost maskovat svůj hlas* [*The Effect of Education on the Ability to Disguise One's Voice*]. Unpublished diploma thesis. Prague: Institute of Phonetics, Faculty of Arts, Charles University.

Wagner, I. & Köster, O. (1999). Perceptual recognition of familiar voices using falsetto as a type of voice disguise. In: *Proceedings of 14th ICPhS*, pp. 1381–1384.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag.

Yumoto, E., Gould, W. J. & Baer, T. (1982). Harmonics-to-Noise Ratio as an index of the degree of hoarseness. *Journal of the Acoustical Society of America*, 71, pp. 1544–1549.

RESUMÉ

Srovnávání mluvčích ve forenzně fonetické praxi spočívá ve vyhodnocování podobnosti hlasů v cílových nahrávkách. Tato úloha může být ztížena – a někdy dokonce i znemožněna – snahou mluvčích měnit svůj hlas a změnit tak svou identitu. Cílem této studie je za prvé zmapovat strategie, které mluvčí při maskování hlasu používají; vycházíme přitom z databáze obecné češtiny, která obsahuje 100 mužských mluvčích a která byla sestavena jako referenční databáze pro forenzní účely. V souladu se zjištěními z jiných jazyků byly nejčastěji využívány změny střední hlasové frekvence, tedy celkové polohy hlasu, druhým nejčastějším způsobem maskování byly fonační modifikace, tedy změny kvality hlasu. K dalším strategiím patří změny rezonančních charakteristik hlasu, melodické nebo temporální změny. Součástí výzkumu bylo i posouzení přirozenosti maskování a náročnosti rozpoznání mluvčího při srovnání původní a maskované nahrávky. U 15 mluvčích, u nichž byla náročnost rozpoznání vyhodnocena jako poměrně vysoká, byla provedena i akustická analýza několika parametrů.

Alžběta Růžičková, Radek Skarnitzl
Institute of Phonetics
Faculty of Arts, Charles University
ruzickovaalzbeta@seznam.cz

TEMPORAL VARIABILITY OF FUNDAMENTAL FREQUENCY CONTOURS

ROBIN HRUŠKA AND TOMÁŠ BOŘIL

ABSTRACT

Intonation is one of the means of performing a speech style. Thus, observing pitch variation in an utterance may be a clue to identifying speech style. We design a cumulative slope (CS) index based upon the amount of pitch variation in a measured F0 contour and the duration of that contour. The more pitch changes there are and the greater their frequency range is, the greater the CS index is. This is confirmed by an experiment we conduct: the CS index of utterances with expressive intonation is higher than that of utterances with neutral intonation, and for utterances with neutral intonation the CS index is higher than for utterances with monotonous or flat intonation. However, as there is a great variability between speakers, the CS index as defined currently, cannot be used to universally differentiate between the styles. Results obtained using automatic voice activity detection (VAD) are close to those obtained with manual VAD and thus the extraction of CS index can be reliably automatized.

Key words: fundamental frequency, melody of speech, stylization, variability of pitch contours

1. Introduction

Intonation is a prosodic feature intrinsic to any spoken text. In a broad sense, intonation is a complex function of speech melody, dynamics and rhythm. In this paper, we focus solely on the melodic component since it is the most salient one and we take a narrower look at intonation as perceptually relevant movement of voice pitch, regardless of its linguistic motivation.

Observing intonation is equally as important as observing the segmental properties of speech, since in everyday communication it carries valuable linguistic and non-linguistic information. In tone languages, intonation has a lexical function and is used to differentiate between word meanings. In many languages it serves a grammatical purpose and helps to distinguish whether an utterance is a question or a statement. It also helps to build the syntactic structure of spoken text. Intonation carries pragmatic information; speakers use it to express their moods, emotions and attitudes. It is used to deliberately

https://doi.org/10.14712/24646830.2017.31

accentuate certain words or syllables. It also has an indexical function; it gives us non-linguistic cues about the speaker, his/her age, gender or social status. (Volín & Bořil, 2013)

In analysing the intonation of an utterance, we usually look for pitch events – movements and patterns that are relevant for the perception of intonation. It is still hard to determine what is and what is not a pitch event and there are different approaches to this issue. A pitch event may be represented by a turning point (a peak, a valley or an inflection point) in the pitch contour, by a simple rising, falling or level pitch movement or by a compound pitch movement such as a rise-fall or fall-rise. We can either observe individual pitch events and patterns and their alignment in relation to the segmental layer, or we can observe overall characteristics of the intonational contour, such as pitch range or the rate at which pitch movements occur. While the former is suitable for examining the intonational features that have a linguistic function (such as lexical tones, pitch accents or boundary tones), the latter may be a source of more abstract linguistic or even non-linguistic information. For example, flat intonation with only slight changes may be a sign of a speaker's boredom or lack of interest, while intonation curves with dramatic rises and falls may indicate the speaker's excitement. In Czech, expressive utterances are often characterised with a wider intonation range and more dramatic pitch changes. Emotional utterances frequently use rising pitch movements, or upward transposition of the nuclear accent (or melodeme) (Palková, 1997).

2. F0 contour, stylization and modelling

Intonation melody is a psychoacoustic variable that cannot be measured instrumentally, but we can measure its acoustic correlate, the contour of voice fundamental frequency, or F0 contour for short. The raw F0 contour extracted from an utterance (e.g. using the pitch extraction algorithm in *Praat* software (Boersma & Weenink, 2016)) includes macrointonation (perceptually relevant pitch movements), but also microintonation (pitch fluctuations caused by segmental properties of the utterance). To filter out the perceptually irrelevant information, we use pitch contour *stylization*.

Stylization is a process that tries to simulate human perception of intonation by smoothing and simplifying the F0 contour so it can be interpreted as a sequence of basic elements such as pitch rises, falls or peaks and valleys, whose properties can be quantified by a set of parameters. The core principle of stylization is *perceptual equality*, i.e. the requirement that an utterance resynthesized with a stylized F0 contour should be perceptually indistinguishable from the same utterance resynthesized with the original contour (Hermes, 2006).

Many stylization algorithms have been proposed so far, ranging from simple, purely statistical ones to those that incorporate advanced perceptual and psychoacoustic phenomena. MoMel (Hirst & Espesser, 1993) is one of the former. It uses quadratic regression to find target points in the contour (in most cases, the points identify with local maxima, minima and inflection points) and then interpolates between them using quadratic splines. The Fujisaki model (Fujisaki & Ohno, 1997) presents a different approach in that it treats the F0 contour as a superposition of "phrase components" and "accent (or tone) components". The phrase component layer is modelled by low-pass filtering weighted impulses (“phrase commands”) and the accent components are modelled by filtering square pulses (“accent commands”) with variable amplitude and duration. The *tonal perception model* (Mertens & d’Alessandro, 1995) assumes that we do not perceive intonation as a continuous contour, but rather as a sequence of syllabic tones. The algorithm detects syllabic nuclei, approximates their pitch with static (level) or dynamic (rise/fall) tones and linearly interpolates the contour outside the nuclei.

In this paper we adopt the tonal perception model for several reasons. Compared to the formerly mentioned stylizations, this one is not just based on statistical smoothing, rather it builds on experimental findings in speech pitch perception. It uses *glissando threshold* to determine whether a pitch change in a syllable nucleus is salient enough to be perceived as a dynamic tone and *differential glissando threshold* to determine whether the change of pitch movement rate is salient enough to be perceived as a compound tone. Another advantage is that the algorithm has been implemented in the intonation analysis software *Prosogram* (Mertens, 2004). It is freely available for download as a set of Praat scripts with a graphic interface.

3. Method and research question

The rate and intensity of pitch events may be essential in deciding whether we perceive speech as monotonous or variable. We suppose that speech with frequently occurring and prominent pitch changes will be generally perceived as expressive and variable. Speech with few melodic events that use a small pitch range will be regarded as rather monotonous. In theory, the simplest monotonous intonation is stylized by one slightly declining straight line. The more variable intonation is, the more pitch events appear in the contour and the more it deviates from a straight line. In effect, more turning points and line segments are needed for a reliable stylization.

Suppose an F0 contour that has been stylized as pitch points interpolated with linear segments. Hruška (2016) defines a *cumulative slope index* (in semitones per second) so that

$$CumSlope = \frac{1}{T_{tot}} \sum_{n=2}^{N} |f(n) - f(n-1)|$$

where T_{tot} is the total duration of voice activity (in seconds), N is the number of discrete points in the pitch contour and $f(n)$ is the frequency in semitones of the n-th point.

Frequency in semitones (ST) is

$$f_{ST} = \frac{12\, log(f_{Hz}/100)}{log(2)}$$

where f_{HZ} is frequency in hertz and 100 is a reference value[1] in hertz for 0 ST.

The CS index is merely the sum of absolute frequency differences between subsequent pitch points divided by the duration of the measured speech segment. Consequently, variable and expressive utterances should show higher CS index values than monotonous utterances.

[1] As the cumulative slope index reflects solely frequency differences, the reference value has no impact and any other value would lead to exactly equal results.

4. Experiments

To test how the proposed index captures/reflects the variability of F0 contours, we conducted two experiments. The data set for Experiment 1 (presented in Hruška (2016)) was taken from the Mini-Dialogue part of the Prague Phonetic Corpus. These scripted dialogues in Czech were performed by students of philological programmes at Charles University. For this study, we randomly chose 5 males and 15 females from the corpus. Each of 10 dialogue lines (see Table 1) was read and acted in neutral style by every speaker, 7 utterances were removed (faulty, omitted or inserted words by the speaker) leading to 193 utterances in total. These recordings were manually segmented, allowing for a precise measurement of duration of voice activity. The purpose of the experiment was to test whether the CS index would be stable across subjects and utterances in the case of similar reading style and to explore its variability.

Experiment 2 was designed to study the impact of neutral / flat / expressive style on the CS index and to test whether a fully automatic voice activity detection (VAD) algorithm could be used to substitute the manual segmentation process without significantly altering the results. Hence, two voice activity detection (VAD) methods were compared. First, fully automatic VAD algorithm was used (Sohn et al., 1999) and second, voice activity segments were marked manually. In both methods, for a pause to be labelled as a voiced-inactive segment, the minimum of 100ms was chosen. Text of three mini-dialogues from the Prague Phonetic Corpus were chosen (15 turns, see Table 2) but completely new recordings were taken with 3 males and 2 females ranging from 23 to 32 years of age (no speaker of Experiment 1 took part in Experiment 2). To test the capability of the CS index to distinguish among different dynamics of F0 contours, the speakers were instructed to perform the dialogues in three styles: a) neutral acting, b) monotonous (flat intonation), and c) with expressively dynamic intonation.

Both Experiment 1 and 2 were recorded in a sound-treated studio. Experiment 1 recordings have 32 kHz sample rate with 16bit PCM, Experiment 2 recordings are sampled with 48 kHz and 16bit PCM. The difference in sample rate should not have any impact on the following analyses as the only concern is about VAD and F0 extraction.

F0 contours were extracted and stylized in Prosogram v2.13 (Mertens, 2004) for Praat (Boersma & Weenink, 2016) with the following settings: *Task – Calculate intermediate files, Segmentation method – Automatic: acoustic syllables*. Then, stylized contours were processed using rPraat (Bořil & Skarnitzl, 2016). Frequency values were converted to semitones and the CS index was calculated for each utterance among all voice-active segments.

Table 1. Experiment 1, 10 utterances.

ID	Text
H1a_5	To bych se nebál.
H2d_1	Máte nějakou představu, jak to bude probíhat?
H3a_1	Až začnou o tom transportu, nastražíš uši.
H3b_5	Na internetu to není?
H3c_1	Trochu se bojím, že je tím rozčílíte.
H4b_1	Hodilo by se trochu víc informací.
H5c_1	Takže všechno to teď závisí na vás.
H5d_5	A ze vzduchu nic vidět není?
H6c_5	A právě s tím nikdo nepočítá.
H6d_1	Jen abyste měli dost peněz, až to přijde.

Table 2. Experiment 2, 15 utterances.

ID	Text
H1b_1	Začni opatrně. A o další spolupráci bych se nezmiňoval.
H1b_2	To by mě ani nenapadlo.
H1b_3	Řekneš jim, co si myslíš?
H1b_4	Nevím, snad nebudu muset.
H1b_5	Dobře, nezapomeň, hlavní heslo: opatrnost.
H1c_1	Nejlepší bude, když zatroubíte a počkáte, až vylezou.
H1c_2	Hmm, no a potom?
H1c_3	Řeknete jim, co si myslíte.
H1c_4	A vy na nás počkáte?
H1c_5	Jasně, ani se nehneme z místa.
H2a_1	Takže se posadíš a budeš se usmívat. Žádnou paniku.
H2a_2	Jasně. A co mám dělat, až přijde ten jejich šéf?
H2a_3	Zeptáš se, co bude dál.
H2a_4	To je všechno? Nemám mu říct, že už to víme?
H2a_5	Ne, nic nevíš. Ani slovo.

5. Results

Results of Experiment 1 are shown in Fig. 1. Fig. 1a shows the distribution of cumulative slope index for each dialogue turn across all speakers. Most values are spread in the interval from 5 to 15 ST/s with no sign of any anomalous (outlier) trend. A few turns seem to be slightly higher (H3a1, H3c_1, and H4b_1) with more than 75% of values larger than 10, which may be due to apparent pragmatic element in their meaning which might have led to the speakes's more dynamic acting. Fig. 1b presents the distribution of CS index of all dialogue turns grouped by speaker. Again, the variability disallows to observe any unique trend, although speaker F9 reaches consistently mildly higher values and speaker F12 seems to remain more often in lower values. These results are not surprising as the conditions of the experiment naturally lead to a neutral acting style which can explain the similar behaviour of intonation dynamics of utterances.

Experiment 2 is presented in Fig. 2 and 3. Figure 2 compares automatic and manual VAD and although manual VAD is certainly more precise (and significantly more time-consuming), the results are very similar. As expected, expressive style leads to higher values of CS index (see Fig. 2a) and monotonous (flat) style produces generally lower values. A detailed look at the dialogue turns (Fig. 2b) shows the same overall trend, although due to evident variability across dialogue turns one cannot determine a threshold that would reliably separate the three styles.

A closer look at individual speakers with manual VAD (see Fig. 3) implies that some speakers performed the three styles of recordings very diversely. On the other hand, female F2 shows very similar range of values for both the neutral and the expressive style. After listening to a few of her utterances, the reason is apparent. To achieve the expressive style, she used other means than intonation such as variation of whisper and shout, vocal timbre or dynamic rhythm changes. As the CS index is meant to measure variability of the F0 contour, this result matches the expectations.

We used R (R Core Team, 2017) and lme4 (Bates, Maechler & Bolker, 2015) to perform a linear mixed effects analysis of the relationship between CS index and style. As fixed effects, we entered style and VAD method into the model (the interaction term was tested using likelihood ratio test but was insignificant with $p = 0.580 >= 0.05$). As random effects, we had intercepts for subjects and utterances, as well as by-subject and by-utterance random slopes for the effect of both style and VAD method. P-values were obtained by likelihood ratio tests of the full model with the effect in question against the model without the effect in question.

Style affected CS index ($p = 0.008 < 0.05$), lowering it by about 5.9 ST/s ± 1.2 standard errors (flat style) and rising it by about 4.7 ST/s ± 1.6 standard errors (expressive style). VAD method also affected CS index ($p = 0.012 < 0.05$), manual VAD rised it by about 1.4 ± 0.5 standard errors in comparison to the automatic VAD. Although statistically significant, this value is low in absolute terms in comparison to the effect of style, and thus the practical impact of the VAD method to the overall use of CS index value to distinguish among styles is not notable.

Detailed look at model coefficients showed that the manual VAD method impact is very consistent across utterances (min. 1.16, max. 1.68 ST/s increase) and also across subjects (min. 0.95, max. 1.72 ST/s). Also, the effect of styles is very consistent across utterances. Although there are differences of the effect of styles across subjects (in three male subjects the differences among styles are more distinct than in two females), the sample size is too small to make any generalization in this sense. Even though each subject had different range of style dynamics, the direction of CS index difference among styles was consistent across all five subjects.

Figure 1. Experiment 1, cumulative slope index of all utterances (a) by utterance, (b) by speaker.

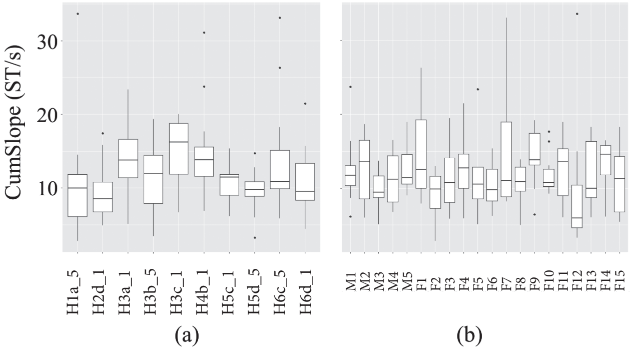


Figure 2. Experiment 2, cumulative slope index for the three intonation styles: neutral, flat, and expressive. Automatic and manual methods of voice activity detection are compared. (a) depicts summary plot of all utterances by all speakers, (b) displays identical data sorted by utterances, i.e. each colour group consists of 15 boxes, each corresponding to one utterance (see Table 2) performed by all speakers.

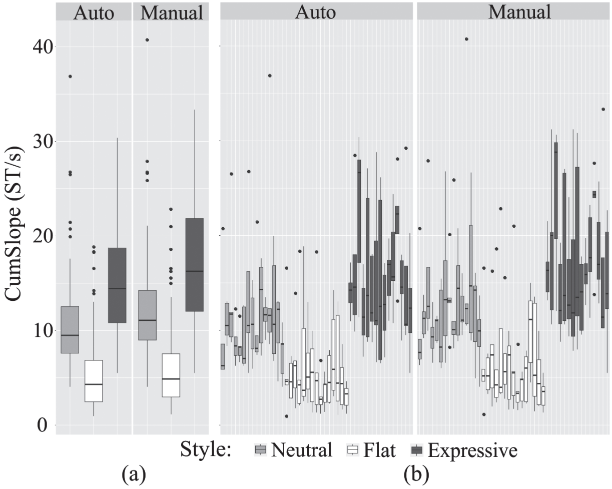
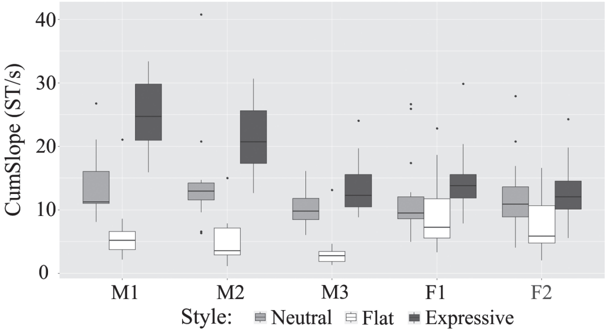


Figure 3. Experiment 2, cumulative slope index of all utterances by 3 male (M1, M2, M3) and 2 female (F1, F2) speakers in three style scenarios, manual voice activity detection only.



6. Discussion and conclusions

In Experiment 1, the speakers did not focus specifically on their acting style. Moreover, they were students of similar age, education and social circumstances. Thus, the cumulative slope values are obviously variable, but no trend can be observed. In Experiment 2, speakers were instructed to stylize their speech with expressive, neutral and flat intonation and the values differ between the individual styles quite consistently across the speakers except for F3, who did not use intonation to achieve the expressive style, but rather used other means, such as variation of articulation rate or vocal timbre. In these cases, the CS index values approach those of neutral style utterances, which is an expected result. However, similarly to Laan (1996), who compared the prosodic properties of read and spontaneous speech, we found that F0 variation alone could not reliably discriminate between the three speech styles.

When we compare the results obtained through automatic voice detection to those obtained through manual voice detection we see the values are fairly similar. Even though the automatic VAD is not as precise, it is consistent in the errors it produces and significantly less time-consuming since it saves manual work. The errors mostly occurred at voiceless plosives or when there were audible breaths or mouth clicks in the recording. They led to longer voice activity having been detected and thus slightly lower CS index values. Nevertheless, this behaviour is consistent across all the styles and speakers and it does not skew the relative differences between them. Thus, CS index can be measured and computed in a fully automatic manner with no human interference required.

The CS index is a simple sum of absolute values of pitch changes divided by the duration of the measured segment. It does not provide information about the actual slope or steepness of the F0 contour, rather it shows how much the contour varies in a given speech segment (in our case a dialogue turn). There is no absolute scale that could be used as a reference for the measured values, since the index seems to depend on the speaker and on the utterance. As it is defined now, it can only be used to compare speech segments with/relative to each other, e.g. two utterances by the same speaker. Furthermore, the CS index does not reflect the shape of the measured contour. A consistently rising intonation may produce the same value as a consistently falling one and no difference is made between a pitch movement at the beginning and at the end of an utterance. However, the CS index was designed to be simple and suitable for fully automatic processing of large quantities of data.

In future research, it would be useful to test a few variants of the CS index, such as instead of dividing the sum of pitch differences by the total duration, dividing it by the number of syllables in the measured segment, obtaining the index value in semitones per syllable. This method would still allow for fully automatic implementation since syllabic nucleus detection is already involved in the Prosogram stylization we use. Another experiment could measure the CS index using a sliding window to see how it changes over time and to test whether it can be used in real-time signal processing, e.g. for monitoring actual mood of a client during a phone call with an operator.

ACKNOWLEDGEMENTS

This research was supported by the Czech Science Foundation project No. GA16-19975S "Age-related changes in acoustic characteristics of adult speech".

REFERENCES

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1–48.

Boersma, P. & Weenink, D. (2016). *Praat: doing phonetics by computer* [Computer program]. Version 6.0.14, retrieved 11 February 2016 from http://www.praat.org/.

Bořil, T. & Skarnitzl, R. (2016). Tools rPraat and mPraat. In: P. Sojka, A. Horák, I. Kopeček & K. Pala (Eds.), *Text, Speech, and Dialogue*, 367–374. Berlin: Springer International Publishing.

Fujisaki, H. & Ohno, S. (1997). Comparison and assessment of models in the study of fundamental frequency contours of speech. In: *INT – 1997*, 131–134.

Hermes, D. J. (2006). Stylization of pitch contours. In: S. Sudhoff et al. (Eds.), *Methods in empirical prosody research*, 29–61. Berlin: Walter de Gruyter.

Hirst, D. & Espesser, R. (1993). Automatic modelling of fundamental frequency using a quadratic spline function. In: *Travaux de l'Institut de Phonétique d'Aix*, Vol. 15, 75–85.

Hruška, R. (2016). *Properties of fundamental frequency contours in segmental contexts*. Unpublished bachelor thesis. Prague: Institute of Phonetics, Faculty of Arts, Charles University.

Laan, G. P. M. (1996). *The Contribution of Intonation, Segmental Durations, and Spectral Features to the Perception of a Spontaneous and a Read Speaking Style*. Amsterdam: Institute of Phonetic Sciences, University of Amsterdam.

Mertens, P. & d'Alessandro, C. (1995). Pitch contour stylization using a tonal perception model. In: *Proceedings of 13th International Congress of Phonetic Sciences*, Vol. 4, 228–231.

Mertens, P. (2004). The Prosogram: Semi-Automatic Transcription of Prosody based on a Tonal Perception Model. In: B. Bel & I. Marlien (Eds.), *Proceedings of Speech Prosody 2004*, Nara (Japan): ISCA.

Palková, Z. (1997). Modelling intonation in Czech: Neutral vs. marked TTS F0-patterns. In: *Intonation: Theory, Models, and Applications*. Athens, Greece.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Sohn, J., Kim, N. S. & Sung, W. (1999). A statistical model-based voice activity detection. *IEEE Signal Processing Lett.*, 6 (1), 1–3.

Volín, J. & Bořil, T. (2013). General and speaker-specific properties of F0 contours in short utterances. *AUC Philologia 1/2014, Phonetica Pragensia XIII*, 9–20.

RESUMÉ

Intonace zásadně ovlivňuje vyznění i celkový význam řeči a při různých mluvních stylech může mít rozličné průběhy. Z toho důvodu je zajímavé sledovat variabilitu časového vývoje intonační výšky jednotlivých promluv. Navrhli jsme ukazatel kumulativní strmosti, tzv. cumulative slope (CS) index, který jedním souhrnným číslem vyhodnocuje míru proměnlivosti kontury základní frekvence řečového signálu (F0) vzhledem k celkovému trvání analyzovaného úseku.

Aby tento ukazatel skutečně pracoval s průběhem vnímané výšky intonace, rozhodli jsme se jej použít na tzv. stylizované kontury F0, které byly vypočteny algoritmem modelujícím percepční dopad vývoje základní frekvence řeči. Předpokládáme, že intonačně dynamičtější promluvy obdrží vyšší hodnotu CS indexu, zatímco monotónnější řeč bude ohodnocena číslem nižším.

V rámci dvou provedených experimentů jsme zkoumali chování hodnot CS indexu. Nejdříve jsme analyzovali nahrávky předem připravených krátkých dialogů, u kterých jsme očekávali podobnou míru dynamičnosti intonace, jelikož se mluvčí adaptovali na obdobný mluvní styl, příhodný pro tento typ úlohy. Obdržené výsledky CS indexu skutečně odhalily, že vzhledem k přirozené variabilitě hodnot ukazatele nebyly výrazné rozdíly mezi jednotlivými mluvčími. Ve druhém experimentu byli mluvčí explicitně instruováni k tomu, aby dialogy namlouvali postupně několika styly, a sice neutrální intonací, expresivní intonací a monotónní, resp. plochou intonací. V této úloze dosahoval CS index v jednotlivých stylech skutečně výrazně rozdílných hodnot. Pokud však byla expresivnost promluvy dosažena jinou cestou než výraznými změnami výšky intonace, tedy například dramatickým střídáním mluvního tempa a přechodem mezi šepotem a hlasitou řečí, CS index nabýval hodnot v rozsahu typickém pro neutrální styl. Takové chování považujeme za správné, protože tento ukazatel byl navržen pro vyhodnocování jedné konkrétní složky intonace a pro měření tempa řeči nebo stylu fonace mohou být použity jiné ukazatele.

Výstupy obou experimentů napovídají, že CS index je do jisté míry nezávislý na mluvčím a odráží míru dynamičnosti průběhu výšky v rámci intonace, která souvisí i s mluvním stylem. Vzhledem ke srovnatelným výsledkům v případě použití automatické detekce řečové aktivity s detekcí manuální je možné proces výpočtu CS indexu spolehlivě plně automatizovat.

Robin Hruška, Tomáš Bořil
Institute of Phonetics
Faculty of Arts, Charles University
robinhruska@email.cz

ACOUSTIC CORRELATES OF PROSODIC DIMENSIONS IN YOUNGER AND OLDER SPEAKERS OF CZECH

JAN VOLÍN AND TOMÁŠ BOŘIL

ABSTRACT

The present study reports phonetic data applicable for diagnostic purposes in voice related pathologies. However, apart from purely physiological concern, linguistic considerations are also acknowledged since the speech material consists of a continuous spoken text. Three age groups of speakers were recorded (young, middle-aged and old adults), each represented by 15 men and 15 women (*n* = 90). Several measures of fundamental frequency, together with variation in intensity and speech tempo were captured. An appreciably innovative metric, Cumulative Slope Index (CSI), was successfully employed to capture F0 variability in utterances. The results confirm differences between the age groups, but also between men and women, and contribute the normative mapping of the Czech population.

Key words: intonation, aging, diagnostics, articulation rate, fundamental frequency

1. Introduction

The speech modulation in the domains of the fundamental frequency, amplitude, timing and spectral setting has been shown throughout decades of linguistic research to encode multiple meanings in communication. The perceptual correlate of this acoustic aggregate of four dimensions is commonly termed *speech prosody* (or intonation in the broad sense of the word) and has been acknowledged to participate significantly in the sound structure of speech or sound patterns of languages (for detailed accounts see, e.g., Bollinger, 1978; Gussenhoven, 2004; Ladd, 2008; Büring, 2016). Various functions of prosodic events are classified into categories, such as *lexical* (phonemic function in the phonological composition of a single word), *grammatical* (signalling types of sentences and indicating their constituents of lower order), *affective* (revealing attitudes and interpersonal stances, moods, emotions), *pragmatic and discourse* (guiding the attention of the recipient and managing his consequential behaviour) and *indexical* (indicating the membership of the speaker in various social groups). It is the last one that relates most directly to the topic of our present study.

https://doi.org/10.14712/24646830.2017.32

Language communities can be described in terms of groupings of the language users who share certain patterns of speech behaviour. These groups can be determined, for instance, by geographical region, education, socioeconomic class, gender or age. Each of such criteria provides interesting and eventually applicable information about the speakers' habits both in the area of speech production and speech perception. The age criterion, for instance, may inform the fields of cognitive and developmental psychology, language and speech acquisition, but also medical disciplines. The latter motivated our research. Our principal question was whether there is a possibility to capture differences between younger and older adults whose mother tongue is Czech, but also what the typical values describing Czech speakers are. In stipulating medical diagnoses, the parameters of speech can be often useful as long as the standards for healthy population are known. In other words, if normal age-related changes are reliably described, pathological conditions that affect voice can be identified.

The interest in the influence of age on voice performance can be traced back across centuries. R. Baken cites Shakespeare who commented on the voice of an elderly man as follows: "… turning again toward childish treble, pipes and whistles in his sound" (Shakespeare, *As You Like It*, Act 2, Scene 7, quoted in Baken, 2005). Yet our attention is attracted quite naturally by less poetic and more systematic approach to the issue: the empirical research of the modern era, i.e., more or less the twentieth century. Here, again, the evidence of keen interest can be observed, but, in addition to that, the methodologically plausible routines dating back many decades (e.g., Bach, Lederer & Dinolt, 1941; Macklin & Macklin, 1942; Birren, 1956 or Mysak, 1959). However, despite the achievements of the past, the available literature also documents that the search for new data and new models is inevitable: R. Baken, one of the leading researchers in the field, expanded and reinterpreted some of his earlier views (cf. Baken, 1987 and Baken, 2005), or S. Linville pointed out serious methodological problems in earlier measurements (Linville, 2000: 364) In addition to that, the research in prosody shows that individual languages may impose specific constraints on physiological mechanisms. Pitch range, for instance, is systematically narrower in Czech than in English in comparable spoken texts (Volín, Poesová & Weingartová, 2015). Therefore, the studies that were carried out abroad and mapped certain important trends cannot be just translated for domestic purposes. It is necessary to investigate the population norms across the linguistically diverse communities.

The age of a speaker has its "signature" in his or her speech. It is known that listeners can differentiate between older and younger voices with remarkable confidence. In research where respondents listened to continuous texts (which is the material of the present study as well), the correlations between perceived and actual age reached the magnitude of about 0.9, which signals a very strong deterministic link (Shipp & Hollien, 1969; Ryan & Capadano, 1978; Hartman, 1979). Hartman also found that female listeners could assess the age of the speaker better than male judges. Yet, if asked about characteristics of older voices, the lay listeners do not necessarily know what really guides their decisions. Certain controversies can be noted in a table compiled by Linville (2000: 361). The listeners, for instance, claim, that elderly persons speak with lower voices regardless their gender. Empirical evidence indicates that this is true only for women, whereas for men the trend is usually ascertained as inverse. Similarly, it is generally believed that aging causes vocal tremor (ibid.). However, many studies have demonstrated that instability in phonation is linked to physical health more strongly than to age. Linville actually suggests that larger frequency and amplitude variation might be better predictors of aging than small fluctuations known as tremor (also jitter and shimmer), not only but also because of methodological problems with measurements (Linville, 2000: 364).

The question arises then as to what methodology is suitable and reliable if we want to quantify the differences that humans distinguish instinctively (or rather implicitly). In addition, it would be useful to know what the contributing factors are and which of them are purely physiological, and which phonological, i.e., connected with different linguistic norms used by younger and older speakers. Our present study contributes modestly to this area of inquiry by testing methodology and by producing descriptive parameters for three age groups of Czech population.

2. Method

2.1 Recordings

Three age groups of native speakers of Czech were recruited. We required healthy individuals without neurological ailments or speech/larynx therapy, with sufficient eyesight and hearing typical for the given age (no special treatment or hearing aids). We also ascertained smoking habits of the subjects as smoking is known to have substantial effect on voice – it generally enhances the aging effects (Gilbert & Weismer, 1974; Braun & Rietveld, 1995; Tarafder, Datta & Tariq, 2012). Hence, smokers were not included in the present sample. The age divisions were: 20–39 years of age for young adults, 40–59 years for middle-aged adults, and 60–79 years for old adults. Each group was represented by 30 subjects (15 male + 15 female). The recording took place in a quiet, comfortably furnished office with a short natural reverberation. High-quality microphone was plugged directly to a portable recorder with uncompressed signal capturing, even though for the parameters investigated in our study the high quality of the recording is not as crucial as it would be, for instance, in the case of jitter, shimmer or breathiness measurements.

The respondents were asked to read out a short extract from a book by a well-known Czech author Karel Čapek (retrieved from selected writings published in Čapek, 1983). The text comprising 137 words does not contain any unusual lexical items or syntactic structures and the speakers were given time to get acquainted with it prior the reading. No special instruction concerning the style of reading was provided. For easier manipulation and better analytical insight, the recordings were divided into 12 units (hypothetical breath-groups, see Table 1).

Table 1. The division of the text into ideal breath-groups for easier analysis.

1	I na tom, že člověk si opatří psa, aby nebyl sám, je mnoho pravdy.
2	Pes opravdu nechce být sám.
3	Jen jednou jsem nechal Mindu o samotě v předsíni.
4	Na znamení protestu sežrala všechno, co našla, a bylo jí pak poněkud nedobře.
5	Po druhé jsem ji zavřel do sklepa s tím výsledkem, že rozkousala dveře.
6	Od té doby nezůstala sama ani po jedinou minutu.
7	Když píši, chce, abych si s ní hrál.
8	Když si lehnu, považuje to za znamení, že si mně smí lehnout na prsa a kousat mě do nosu.
9	Přesně o půlnoci s ní musím provádět velkou hru, při níž se s velikým hlukem honíme, koušeme a kutálíme po zemi.
10	Když se uřítí, jde si lehnout.
11	Pak si smím lehnout i já, ovšem s tou podmínkou,
12	že nechám dveře do ložnice otevřené, aby se Mindě nestýskalo.

2.2 Analyses

The sound supervision and measurements were performed predominantly in Praat v6.0.14 which, apart from computations themselves, allows for labelling the sounds (Boersma & Weenink, 2016). Intensity contours were obtained with 10-millisecond time step by cubic interpolation from intensity objects (minimum pitch 50 Hz, time step auto). F0 contours and stylized F0 contours based on the tonal perception model (Mertens & d'Alessandro, 1995) were computed in Prosogram v2.13 (Mertens, 2004) with the following settings: Calculate intermediate data files (no graphics files), Time range all, F0 detection range 0–450 Hz, parameter calculation Full (saved in file), Frame period 0.005 sec, Segmentation method Automatic: acoustic syllables, Thresholds G = 0.16-0.32/T^2 (adaptive), DG = 30, dmin = 0.050. All frequency values were subsequently converted to semitones (ST) with the reference value of 100 Hz.

The processing of the extracted measurements and analyses of the data were performed in R (R Core Team, 2016) with the application of rPraat (Bořil & Skarnitzl, 2016). To focus on F0 contour variations that is not under the speakers' control, we have also created *alternative F0 contours* by subtracting linearly stylized contours from the raw F0 tracks represented by pitchtiers. Apart from the arithmetic mean we also calculated F0 range (from 5th to 95th percentile). One data point in correlation scatterplots and calculations represents a speaker. (We wanted to avoid data inflation resulting from regarding each breath-group a data point.) Also, due to the log-normal distribution of the measured phenomena, we had to perform logarithmic transformation of the data.

Apart from conventional measurements we also measured the Cumulative Slope Index (Hruška, 2016; for detailed evaluation see also Hruška & Bořil, 2017 – this volume). It is a metric that allows for quantification of the amount of variation in contours (e.g., F0 or intensity contours in relation to their duration). We have adapted this measure for speech in that instead of physical time we used the number of syllables as a timing unit. Cumulative Slope Index (CSI) relative to the number of syllables is computed as follows:

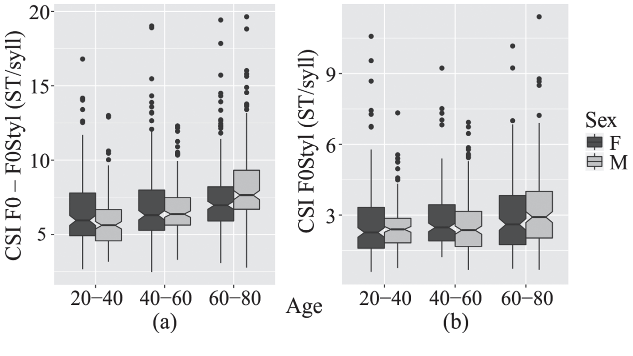
$$CSI = \frac{1}{N_{syll}} \sum_{n=2}^{N} |x(n) - x(n-1)|$$

where N_{syll} is the number of syllables in the utterance, N is the number of discrete points in the analysed contour and $x(n)$ is the value of the n-th point (either in semitones for pitch contours or in decibels for intensity contours).

3. Results

As our primary interest was in the Cumulative Slope Index (CSI), which was an innovative (or less commonly used) metric designed to reflect variation in contours, we report it first. Figure 1 comprises two panels. Panel (a) on the left captures the variation in alternative F0 contours (see above), whereas panel (b) on the right relates to the stylized contours of F0. The alternative contours suggest that with growing age the variation grows as well and the trend is more salient in male than in female production. The stylized contours (panel b) do not display the trend. This means that the crude melodic course which reflects the phonology of the language is not differentiating between older and younger speakers in terms of CSI, while the melodic movements that are outside phonology (we might think of them as being outside the speakers' control) have discriminative power.

Figure 1. Cumulative Slope Index (in semitones per syllable) (a) for the alternative contours (i.e., stylized contour subtracted from the raw F0 track), (b) for the plain stylized contours for three age groups of speakers with gender differentiated. (Notice the difference in the scale of the *y* axis between panels *a* and *b*.)



Variation of F0 above pertains to intonation in the narrow sense of the word. Figure 2 complements the previous account with data that map the tempo and loudness modulations (i.e. speech rate and intensity, respectively). It is obvious that the articulation rate decreases with age. A visible exception is the group of middle-aged males, but considering the confidence intervals, this exception does not seem to be very important.

Cumulative Slope Index for intensity contours seems to grow with age for female speakers, but young males go against this trend in the male sample. It has to be emphasised, though, that the relationship between intensity and loudness is extremely complex so attempts to explain this trend should remain very cautious.

Figure 2. Panel (a) – speech rates in syllables per second and Panel (b) – Cumulative Slope Index for intensity in decibels per syllable for three age groups of speakers with gender differentiated.

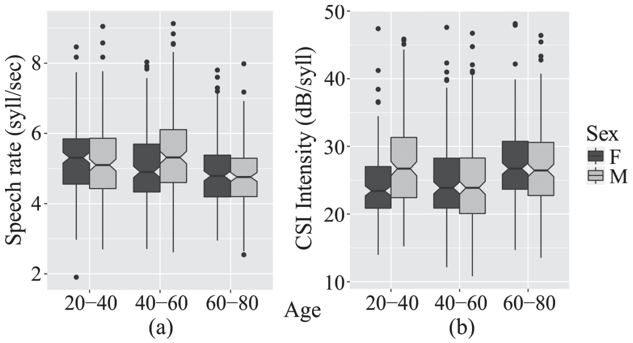


Figure 3. Panel (a) – mean fundamental frequency in semitones (re 100 Hz) and panel (b) – F0 range between the 5th and 95th percentile for three age groups of speakers with gender differentiated.

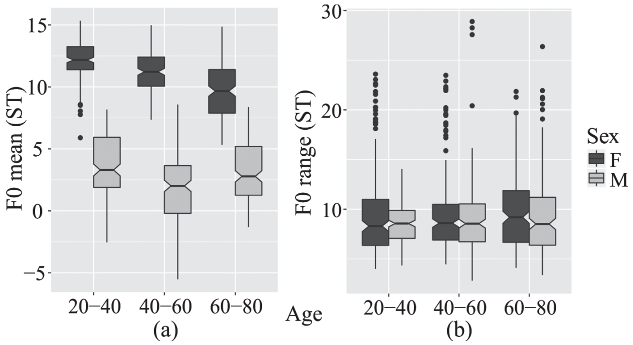


Figure 3 confirms to some extent the antagonistic behaviour of F0 level for male and female population found elsewhere. Women speak with lower voices as they grow older, men produce a scooping trend with the middle-aged group using the lowest values. The question is whether this scoop relates to the intensity variation (Figure 2, panel b), which captures a similar trend. The 90%-range of F0 values did not reveal any age related trend, although the male samples visibly increase the variance in the range values while keeping the median more or less equal. In other words, as to the F0 range, young males produced a compact set of values, while old males displayed a notable within-group dispersion of values.

Figure 4 captures correlations of speech rate with our two measures of contour variation. Both men and women behave in a very similar manner. On the left, the intensity is inversely correlated with the tempo at $r = -0.71$ (Pearson corr. coefficient). This suggests that the faster the speech, the less varied in intensity it is. A similar, but weaker trend is displayed in panel *b* of Figure 4. The alternative F0 contours (i.e., those where stylization is subtracted from the F0 track – see above) also vary less at higher tempos. The Pearson correlation coefficient $r = -0.44$. As already stated, these trends should be only interpreted after further experiments (but see Discussion below).

Figure 4. Scatterplots with trendlines (and 95%-confidence bands): panel (a) – speech rate against variation in intensity, panel (b) – speech rate against variation in F0 (alternative contours). Female data points and trendline are black, male are grey.

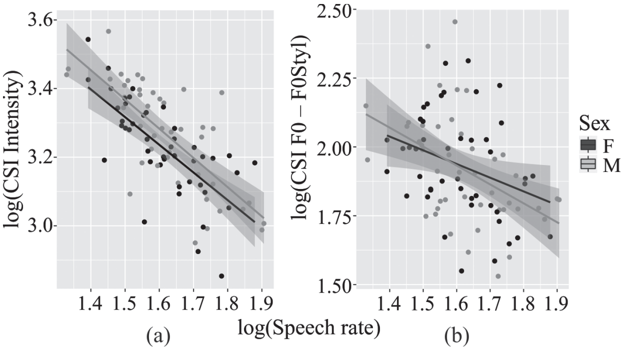
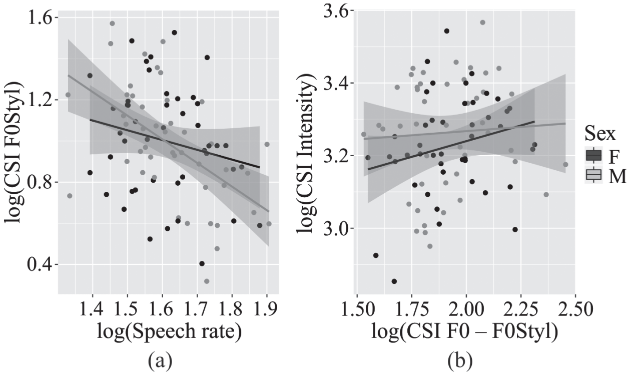


Figure 5. Scatterplots with trendlines (and 95%-confidence bands): panel (a) – speech rate against variation in stylized F0 contours, panel (b) – variation in F0 (alternative contours) against variation in intensity contours.



For the sake of completeness we also correlated speech rate with the stylized F0 contours which are believed to capture basic phonological properties of the intonation (Figure 5, panel *a*). The trend is not very strong: Pearson correlation coefficient $r = – 0.38$. Even weaker is the link between variation in intensity and F0 in alternative contours: Pearson correlation coefficient $r = 0.19$. Here we might speculate about the common denominator, the speech rate, provided it really influences the variation in the two plotted domains.

4. Discussion

Prosodic variations in spoken texts fulfil important communicative tasks, but they also accommodate for individual and group variation. Our task was to map the values typical of selected age groups in Czech population. In general, we confirmed a trend found elsewhere that with aging, men increase their mean F0 (≈ pitch level), while women speak lower (see Ryalls et al., 1994 for overview). In addition, the decrease in articulation rates as a function of age was confirmed in our sample. It has to be stressed, though, that our primary goal was to obtain values typical of Czech population, rather than replicate findings from other studies.

What we find noteworthy is that our material also suggests an increase in variability of F0 and intensity contours with age. On the one hand, this might be a manifestation of the fact that fast articulation rates do not allow for proper prosodic variation: fast speakers make fewer prosodic boundaries and fewer prominences. On the other hand, the informal auditory inspection of the recordings suggested a pragmatic factor: it sounded as if the older speakers enjoyed the performance and wanted to show how lively or engaged their reading can be. This is also confirmed by the behaviour of some of the elderly subjects who wanted to recite poems they learnt by heart at school when they were students. Contrary to that, younger readers displayed mild anxiety not to make reading errors, or to read in a flawless manner. They sounded as if inexperienced with the use of their voice for loud reading. It seems that perceptual testing and development of some more linguistically sensitive metrics will be inevitable if we aspire at solving this and similar dilemmas.

The previous sentence has some bearing on another issue raised in the present study. We noticed that alternative F0 contours capture the age differences while the stylized F0 contours do not. (As explained in the section *Method* above, alternative contours are residual contours after subtraction of stylized contours from the raw extracted F0 tracks.) It is widely believed that stylized contours capture the intonational phonology of a language more clearly than raw contours (see Hermes, 2006 for overview). However, our CSI metric is not sensitive to the alignment of the melodic movements with the syllables of the utterances. Therefore, no definite statement about intonational phonology of younger and older Czech speakers can be made at this stage.

The age as a factor of variation in prosodic features is traditionally considered extralinguistic, and as such has been mostly studied in medical research, recently also by sociolinguistics. "Extralinguistic" is an immensely unfortunate label. Linguists should not light-heartedly relinquish anything that is connected with speech communication for others to study. If they do (and they have done formerly), many interesting facts will be discovered outside linguistics, whilst grammar itself will not guarantee understanding the competences of language users and their speech behaviour.

ACKNOWLEDGEMENT

This research was supported by the Czech Science Foundation project No. GA16-19975S "Age-related changes in acoustic characteristics of adult speech".

REFERENCES

Bach, A. C., Lederer, F. L. & Dinolt, R. (1941). Senile changes in the laryngeal musculature. *Arch. Otolaryng. 34*, pp. 47–56.

Baken, R. J. (2005). The aged voice: A new hypothesis. *Journal of Voice 19/3*, pp. 317–325.

Baken, R. J. (1987). *Clinical Measurement of Speech and Voice*. Boston: Allyn & Bacon.

Birren, J. E. (1956). The significance of age changes in speed of perception and psychomotor skills. In: J. E. Anderson (Ed.) *Psychological Aspects of Aging*. Washington: Amer. Psychol. Association.

Boersma, P. & Weenink, D. (2016). Praat: doing phonetics by computer [Computer program]. Version 6.0.14, retrieved 11 February 2016 from *http://www.praat.org/*.

Bollinger, D. (1978). Intonation across Languages. In: J. Greenberg (Ed.). *Universals of Human Language*, Vol. 2. pp. 471–524. Stanford: Stanford University Press.

Bořil, T. & Skarnitzl, R. (2016). Tools rPraat and mPraat. In P. Sojka, A. Horák, I. Kopeček, & K. Pala (Eds.), *Text, Speech, and Dialogue*. Springer International Publishing, 367–374.

Braun, A. & Rietveld, T. (1995). The influence of smoking habits on perceived age. In: *Proceedings of 13th International Congress of Phonetic Sciences*, Vol. 2, pp. 294–297. Stockholm: IPA.

Büring, D. (2016). *Intonation and Meaning*. Oxford: Oxford University Press.

Čapek, K. (1983). *Spisy: Zahradníkův rok / Měl jsem psa a kočku / Kalendář*. Praha: Československý spisovatel.

Gilbert, H. R. & Weismer, G. G. (1974). The effect of smoking on the speaking fundamental frequency of adult women. *Journal of Psycholinguistic Research 3*. pp. 225–231.

Gussenhoven , C. (2004). *The phonology of tone and intonation*. Cambridge: Cambridge Univ. Press.

Hartman, D. E. (1979).The perceptual identity and characteristics of aging in normal male adult speakers. *Journal of Communication Disorders 12*, pp. 53–61.

Hermes, D. J. (2006). Stylization of pitch contours. In: S. Sudhoff et al. (Eds.) *Methods in empirical prosody research*. Berlin: Walter de Gruyter, pp. 29–61.

Hruška, R. (2016). Properties of fundamental frequency contours in segmental contexts. *Prague: Institute of Phonetics, Charles University, Faculty of Arts, bachelor thesis*.

Hruška, R. & Bořil, T. (2017). Temporal variability of fundamental frequency contours. *AUC – Philologica, Phonetica Pragensia*, pp. 35–44.

Macklin, C. & Macklin, M. (1942). Respiratory system. In: E. V. Cowdry (Ed.) *Problems of aging* (2nd ed., pp. 185–253). Baltimore: Williams & Wilkins.

Mertens, P. & d'Alessandro, C. (1995). Pitch contour stylization using a tonal perception model. In: *Proceedings of 13th International Congress of Phonetic Sciences*, Vol. 4, pp. 228–231.

Mertens, P. (2004). The Prosogram: Semi-Automatic Transcription of Prosody based on a Tonal Perception Model. In: B. Bel & I. Marlien (Eds.) *Proceedings of Speech Prosody 2004*, Nara (Japan): ISCA.

Mysak, E. (1959). Pitch and duration characteristics of older males. *Journal of Speech and Hearing Research 2*, pp. 46–54.

Ladd, D. R. (2008). *Intonational phonology*. 2nd edition. Cambridge: Cambridge University Press.

Linville, S. E. (2000). The aging voice. In: R. D. Kent & M. J. Ball (Eds.) *Voice Quality Measurement*, pp. 361–376.

R Core Team (2016). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*. URL https://www.R-project.org/.

Ryalls, J., Le Dorze, G., Lever, N., Ouellet, L. & Larfeuil, C. (1994). The effects of age and sex on speech intonation and duration for matched statements and questions in French. *JASA 95/4*, pp. 2274–2276.

Ryan, E. B. & Capadano, H. I. (1978). Age perceptions and evaluative reactions toward adult speakers. *Journal of Gerontology 33*, pp. 98–102.

Shipp, T. & Hollien, H. (1969). Perceptions of the aging male voice. *Journal of Speech and Hearing Research 12*, pp. 703–710.

Tarafder, K. H., Datta, P. G. & Tariq, A. (2012). The aging voice. *BSM Medical University Journal 5/1*. pp. 83–86

Volín, J., Poesová, K. & Weingartová, L. (2015). Speech melody properties in English, Czech and Czech English: Reference and interference. *Research in Language 13(1)*, pp. 107–123.

RESUMÉ

Studie přináší sadu referenčních fonetických deskriptorů, které je možno využívat pro diagnostické účely nejen ve foniatrické oblasti, ale i všude tam, kde se patologické změny odrážejí i v hlase pacienta. Ovšem kromě čistě fyziologických aspektů jsou pojednány i aspekty lingvistické, neboť řečovým materiálem je souvislý mluvený text, nikoli izolované jednotky nebo logatomy. Studie pracuje s nahrávkami tří věkových skupin dospělých mluvčích, z nichž každá je reprezentována 15 muži a 15 ženami (tj. celkový počet subjektů = 90). Měření byla prováděna ve třech základních akustických doménách: frekvenční, temporální a intenzitní, a to různými metodami. Jednu z nich, kumulativní index změny (Cumulative Slope Index – CSI) lze považovat za inovativní. Tento index v sebraných datech prokázal svou užitečnost. Celkově potvrdily výsledky studie rozdíly nejen mezi věkovými skupinami, ale i mezi muži a ženami v interakci s věkem. Tyto výsledky jsou příspěvkem ke stanovení chybějících populačních norem pro české jazykové prostředí.

Jan Volín, Tomáš Bořil
Institute of Phonetics
Faculty of Arts, Charles University
jan.volin@ff.cuni.cz

2017 ACTA UNIVERSITATIS CAROLINAE PAG. 55–64
PHILOLOGICA 3 / PHONETICA PRAGENSIA

PITCH RANGE OF INTONATION CONTOURS IN ENGLISH CZECH

JAN VOLÍN, DAMIEN GALEONE AND WESLEY JOHNSON

ABSTRACT

Pitch range is believed to code important information that is indispensable for correct decoding of spoken messages. Previous research found differences in pitch variation across languages like English, French, Bulgarian, Polish, Czech and German. In addition, differences in pitch range of foreign-accented and native speech were found in various types of speech material. In the present study a sample of sixteen English and American men and women produced recordings of spoken texts consisting of eight paragraphs taken from Czech news broadcasts. Manually corrected F0 tracks provided a possibility to extract four measures of F0 distributional dispersion in order to map global intonational habits of Anglophone learners of Czech as a foreign language. The extracted values were compared with reference values from earlier studies. The results in all four measures indicate that foreign accented Czech is spoken with a pitch range that is narrower than that of English and often even narrower than that of native Czech. Considering results of similar, albeit smaller, studies done earlier, we would attribute our findings to implicit uncertainty in the use of the foreign language, rather than to overcompensation.

Key words: speech melody, pitch range, Czech-accented English, English-accented Czech, F0 contours.

1. Introduction

Traditionally, learners of foreign languages aspired at proficiency comparable with that of the native users of the language. Common sense would have it that the educated native speakers' mastery was the appropriate model for a determined language learner. Current teaching insists less on perfection in pronunciation and sets more realistic goals for the majority of L2 learners. The overwhelming trend is to subscribe to the concept of comfortable intelligibility. Although this is a praiseworthy approach, we are far from understanding where exactly intelligibility starts and ends, and, above all, what makes it comfortable. Abercrombie's indication of speech "which can be understood with little or no conscious effort on the part of the listener" (Abercrombie, 1956: 37) sounds very

https://doi.org/10.14712/24646830.2017.33

reasonable but might not be easy to specify in measurable features. In other words, comfortable intelligibility is a reasonable concept in common classrooms of mass education, but will not provide any useful guidance in empirical research unless comfort on various levels of (un)consciousness can be measured.

There is a relatively long tradition in linking intelligibility to segmental phonemic issues. Teachers are often trained to teach pronunciation of minimal pairs like *peak* and *pick*, *light* and *late* or *mouse* and *mouth*. That is by no means wrong, but the belief that such minimal pairs encapsulate the phonetics of the foreign language and that their training is sufficient to acquire nativelike accent is utterly misguided. At the current stage of our knowledge we should presume that any element of pronunciation can contribute to listeners' discomfort (Anderson-Hsieh, Johnson & Koehler, 1992; Derwing, Munro & Wiebe, 1998; Jilka, 2000; Derwing & Rossiter, 2003; Field, 2005; Kang, Rubin & Pickering, 2010) and, therefore, various features of foreignness should be held suspect of disrupting the smooth flow of perceptual processing. All aspects of sound patterning in speech may produce mismatches between the expected forms and the incoming acoustic signal (cf. Grossberg, 2003). These are likely to activate additional cognitive resources and, subsequently, put strain on vital brain capacities like attention or working memory (Van Engen & Peelle, 2014). Therefore, we consider it legitimate to expand our understanding of foreign accentedness into the area of speech prosody. Our concern in the present study is speech melody or intonation in the narrow sense of the word.

From a lay perspective the functions of intonation might seem purely ornamental. This illusion might be caused by the fact that intonation is not easy to access for conscious evaluation. However, intonologists or conversation analysts can provide abundant evidence concerning the importance of melodic events in speech for effective communication. Without clear prominences and phrasal breaks the speech becomes muddled, difficult to follow, and essential pragmatic and affective messages like friendliness or willingness to cooperate might be absent (e.g., Gilbert, 2014).

Volín and Poesová (2016) carried out an experiment in which they measured reaction times as indicators of the ease of cerebral processing on the part of listeners. They used semantically unpredictable utterances recorded by native speakers of English and by Czech learners of English. Test items were also created by hybridizing the F0 tracks: melodies produced by native speakers were implanted on the Czech-produced utterances and vice versa. Thus, there were four conditions for each of the utterances: (1) native English speech with native English intonation; (2) native English speech with Czech English intonation; (3) Czech English speech with native English intonation; (4) Czech English speech with its specific intonation. Conditions 1 and 4 were the original recordings (resynthesized without changes in order to equalize the technical quality of the sound), while conditions 2 and 3 were hybridized. The results showed that this plain swapping of F0 tracks had an impact on reaction times and that unaltered Czech English was the most difficult to process (Volín & Poesová, 2016). Unfortunately, the authors did not analyse the differences between the melodies in detail, but made it clear that one of the noticeable facts was the difference in the span of melodic movements.

Multiple attempts to describe melodic events in speech have produced large numbers of descriptive methods. The latest advances in speech technologies (speech synthesis, automatic speech recognition) seem to suggest that descriptions of speech melodies (i.e., intonation in the narrow sense of the word) can be either cognitively accessible or technologically applicable but not both. Speech recognizers and synthesizers use massive computational algorithms that can produce the desired results but, unfortunately, cannot be turned into explanations of how speech melodies function. On the other hand, statements like "rising tune" or "monotonous melody", which would make sense to most people, are too vague to be used in rigorous intonation modelling. Most linguists, psychologists and language teachers would like to see some sort of a compromise: a descriptive system (probably an open one, without a fixed preconceived inventory) that allows for a clear link between perceptual categories and specific melodic events (definable by specific physical properties).

One of the ways to describe melodic events is to consider the minimal building blocks of the phrase tunes (termed, e.g., pitch accents, tones), the pitch range in which they are produced, and the combinatory and distributional context in which they are used. Of this triad (*pattern – range – phonosyntax*) we will focus on the second aspect. For the purpose of detailed intonological description of local melodic events, Ladd (2008) proposes to consider pitch range a two-dimensional construct, with the dimensions of level and span. Level is known from older terminology as register, but this label is perhaps too easily confused with a long-term voice setting (bass, baritone, mezzosoprano, etc.) rather than a parameter that speakers change even within an utterance (e.g., the prosody of parenthetical clauses).

Nevertheless, our study focuses on more global attributes of speech so the dimension of level will not be investigated. We will address the question raised, for instance, by Hirst and Di Cristo, who wondered if various languages could be spoken with different overall pitch ranges (Hirst & Di Cristo, 1998: 42). At the time of writing their survey of intonation systems, they did not have any data to answer the question. This is because the seemingly simple matter of global pitch range can be meaningfully addressed only if the prerequisite speech samples are collected under comparable conditions and are of sufficient size. Comparable conditions are necessary since there might be differences in pitch range settings across various speaking styles, but also due to the fact that pitch range signals important affective messages that reflect the conditions under which the speech is produced (e.g., Patterson, 2000; Scherer, 2003). A sufficient size of the sample neutralizes the fact that individuals may differ in their habitual pitch range quite substantially. This should be also observed in our current data.

Studies comparing pitch range across languages do exist. Hirst himself, for instance, tested a rather complex set of F0 descriptors and successfully differentiated 10 English speakers from 10 French speakers (Hirst, 2003), yet it seemed that global pitch range did not play any role. He later added 10 Chinese speakers to demonstrate the merits of his method (Hirst, 2013). The report of Keating and Kuo (2012) concerning the difference between English and Chinese is also inconclusive with regard to pitch range, and, importantly, cautions researchers about the influence of the type of speech material. Mennen and her colleagues found that German spoken texts were produced with narrower pitch range than comparable texts in English (Mennen, Schaeffler & Docherty, 2007). Eight authors participated in a project to measure pitch ranges in four languages: Bulgarian, English, German and Polish (Andreeva et al., 2014). They used continuous texts read out by speakers on request. Interestingly, their results did not confirm the difference found by Mennen and colleagues for English and German, but found significant differences between the two Slavic and two Germanic languages. In general, Polish and Bulgarian displayed greater pitch variation than English and German. A potential problem might stem from the fact that the group of authors did not record their own speech material. Instead, they used already existing corpora from which they selected speakers based on undisclosed criteria.

Volín, Poesová and Weingartová (2015) used an extensive sample of data to provide reliable reference values for English and Czech read monologues. They used texts of news bulletins read out by 32 professional newsreaders from national radio stations. Their results will be used in order to put the outcome of the present study in perspective.

It is only natural that the question of the overall pitch variation has entered the field of foreign or second language (in this study we use L2 for both) acquisition. If languages differ in the mean pitch range used, how will this fact influence the accented speech of L2 learners? Jun and Oh asked four learners and two native speakers to read a set of 40 specially constructed sentences. Their objective was to investigate the acquisition of Korean intonation by speakers of American English, placing special emphasis on phrasing. Among other things pitch range produced in foreign-accented Korean was narrower (Jun & Oh, 2000). The possibility of L2 learners using narrower pitch ranges is further corroborated by Lee (2014) who recorded eight Anglophone speakers learning French. Although he only measured phrase-final rises, he found that the learners spoke with narrower pitch range.

2. Method

Sixteen Anglophone speakers of Czech (eight women and eight men) were asked to read out a news bulletin originally broadcast on Czech Radio (a national broadcaster). They were all resident in Prague and spoke Czech at the levels of B1 to C1 of the CEFRL. The length of their residence in the Czech Republic varied from 1 to 20 years but this did not correlate with their L2 proficiency.

The subjects were given a print-out of the text of six paragraphs and were given sufficient time to get acquainted with it. The recordings were made in a sound treated room with a condenser microphone connected directly to a computer sound card. The recordings were saved in an uncompressed format at a 32-kHz sampling frequency using a 16-bit resolution. The spoken text was divided into breath-groups with a constraint on the excessive breathing producing breath-groups too short. Any breath-group shorter than 1.2 sec. was left with the adjacent one (preceding or following considering the prosodic closeness). Each speaker produced about 55 breath-groups with a mean duration of 5.2 sec.

F0 tracks were extracted in the speech analysis software Praat (Boersma & Weenink, 2014) with the autocorrelation algorithm. Individual values were taken in 10-ms steps and the contour was smoothed by a 10-Hz filter. Subsequently, all 872 contours were manually corrected since some of them contained octave jumps, spurious periodicities in voiceless regions or missing F0 values in soft breathy syllables. All contours were also interpolated through the voiceless regions to emulate the human percepts, which are also uninterrupted by the voiceless consonants (cf. Volín & Bartůňková, 2015).

Four correlates of pitch range, which are termed measures of data dispersion in descriptive statistics, were computed. In this study they will be referred to as: variation range (VAR), 80% percentile range (PER), interquartile range (IQR) and standard deviation (SD). Since comparisons of results across various studies are desirable, our understanding of the measures will be explained in more detail.

By *variation range* the span over the values or the distance between the lowest and the highest value (minimum – maximum) is understood. Although this measure is very popular, its disadvantage lies in the fact that it hinges on two extreme measure only, hence it is quite prone to error. We overcame this disadvantage in our study by computing the arithmetic mean of more than fifty breath-group ranges for each speaker. Thus, the range for a given speaker is not dependent on just two values, but on more than a hundred measurements. (Similarly, when a maximum or minimum is mentioned further in the text, it is not the absolute maximum of the given speaker, but the mean maximum averaged across all the breath-groups produced by that speaker).

Percentile range is a general term for distances between specified points in ordered rows of values. Even though, in theory, any two points can be selected, researchers use either memorable or otherwise justifiable numbers. In our study we will report the distance between the 10th and the 90th percentile. This measure is also mentioned as a possibility by Patterson and Ladd (1999) and used by Mennen et al. (2007) and Lee (2014). This 80 % percentile range will be referred to as PER for short. Again, the PER is computed for each breath-group produced by a speaker and the arithmetic mean for that speaker is reported.

Interquartile range (IQR) expresses the distance between the 25th and the 75th percentile in an ordered row of values. In other words, the lowest and the highest quarters of the data are disregarded and the variation range of the medial half of the ordered data is measured. This measure was also used in the above-mentioned study by Andreeva and colleagues (2014). The advantage of this measure is its stability: it is not influenced by extreme values. On the other hand, it can also lack important specific information if the domain of a phonetic feature happens to be outside the range of typical values.

Standard deviation is, in a sense, an improved concept of mean deviation. It also approximates the dispersion of the values around the arithmetic mean but, for the sake of generalizability, it weighs smaller distances from the mean differently from bigger distances, and takes into account the size of the sample from which it is calculated. Its use is widespread although it seems that it is sometimes forgotten that SD is designed primarily for symmetrical data. F0 values are usually asymmetrical – skewed to the right. Unlike ranges, SD values will be reported in Herz (Hz), which, relative to semitones (ST), is an exponential unit. Therefore, male and female results must be presented separately.

3. Results

The important aspect of the present study is the consideration of the reference values of both native Czech and native English. These were provided by Volín, Poesová and Weingartová (2015) and pertain to the same type of spoken texts. Figures 1 and 2 present the reference values together with the means of the range measurements obtained from the current material.

Figure 1. Mean values of three types of F0 range measures: variation range, 10–90 percentile range and interquartile range (see text) for English, Czech and Czech spoken by native speakers of English.

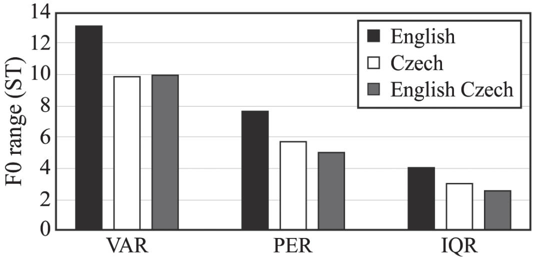
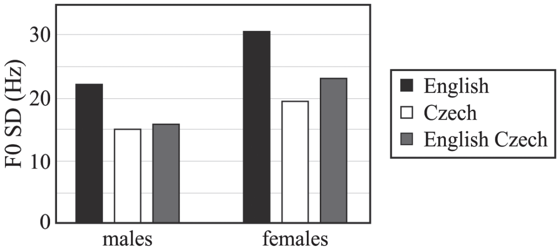


Figure 2. Mean values of F0 standard deviations for male and female speakers of English, Czech and English-accented Czech.



Since all of the differences between the Czech and English reference values were ascertained as highly statistically significant with $p < 0.001$ (Volín, Poesová & Weingartová, 2015), only t-tests for referential values were calculated to see whether the English-accented Czech (E-Cz) differed significantly. With regard to VAR, E-Cz was narrower than native English by 3.7 semitones, which was found highly significant: $t(15) = 7.39$; $p < 0.001$. Contrary to that the difference from native Czech was highly insignificant as it amounted to less than one tenth of a semitone ($p > 0.88$).

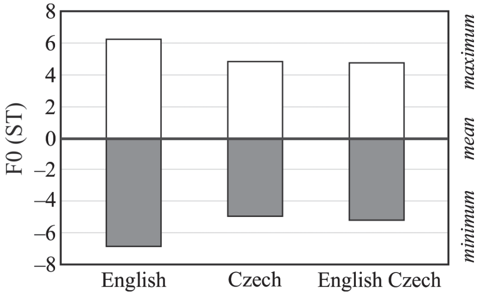
For PER the difference between native English and E-Cz was 2.6 ST and it was highly significant: $t(15) = 8.61$; $p < 0.001$. The difference between native Czech and E-Cz is only 0.68 ST but even this small number reached statistical significance since we were dealing with relatively stable concentrated values: $t(15) = 2.22$; $p < 0.05$.

The IQR produced the difference between native English and E-Cz of 1.5 ST and this was found to be highly significant: $t(15) = 8.93$; $p < 0.001$. Cz-E interquartile range was also narrower than that of native Czech by about 0.5 ST. Even this result reached statistical significance due to concentration of the values: $t(15) = 2.88$; $p < 0.05$.

Statistical significance of the standard deviation metric was calculated separately for men and women, i.e., with only 7 degrees of freedom (8 men and 8 women). Unlike Mennen et al. (2007) we used a two-tailed test, which is more rigorous. The lower SD for E-Cz compared with native English reached significance for both men and women: $t(7) = 5.03$; $p < 0.05$ and $t(7) = 3.27$; $p < 0.05$, respectively. As displayed in Figure 2, SD for E-Cz was higher in comparison with native Czech. This difference was found significant for women: Cz $t(7) = 2.48$; $p < 0.05$, but not for men: $t(7) = 0.34$; $p > 0.74$.

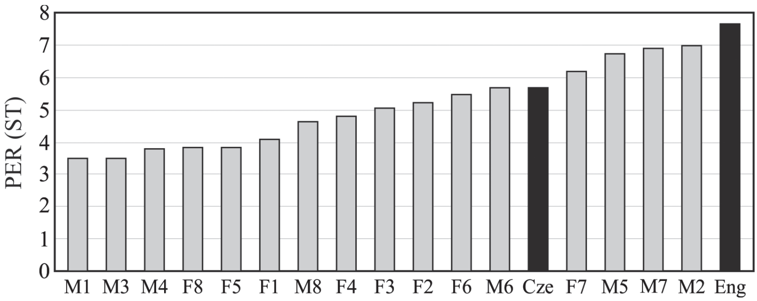
Apart from ranges that express the dispersion of F0 values we were also interested in the position of the lowest and highest values relative to the arithmetic mean. Figure 3 captures the situation. Clearly, there is no difference in the symmetry or asymmetry of the values: under each of the three conditions the speakers depart slightly further down from the mean than up. However, the magnitude of this difference extends to only fractions of a semitone. The limits of PER and IQR offered a very similar picture (on a smaller scale) and SD stretches to an equal distance up and down from the mean by definition.

Figure 3. Mean maxima and minima of F0 values relative to the arithmetic mean normalized to zero. The three columns represent native English, native Czech and English-accented Czech.



Another important question concerns the behaviour of individuals within the studied group. This is because mean group values are more reliable and applicable in predictions if individuals do not depart to far from the mean. Figure 4 presents PER values for each of the speakers in our sample together with the reference values for native English and Czech. It is the decomposition of the middle triplet of values from Figure 1.

Figure 4. Mean 80% percentile range (PER) for individual speakers from the sample (grey columns) together with the reference values for native Czech (Cze) and native English (Eng) represented by the black columns. M = male speaker, F = female speaker; numbering of the speakers bears no relevance to the results.



Obviously, what can be stated about the group does not necessarily hold for all the individuals. Four of the sixteen speakers produced values that fell between the Czech and English reference values. They would comply with the notion of interlanguage. However, majority of the speakers produced values below both referential points.

Figure 4 also reveals that there is no clear division between male and female speakers. The three narrowest ranges were produced by men, but so were the three broadest ranges. Similarly, we did not find any connection between the produced ranges and the length of residence of the speakers in the Czech Republic.

4. Discussion

Pitch range in spoken texts is not random or chaotic. It is used systematically and apart from universal functions (e.g., to signal affective arousal) it also seems to display language specific features. The main obstacle in describing the phenomenon clearly and comprehensively is the lack of methods that would be feasible to employ (given the underfinancing of phonetic research), and satisfactorily sensitive to important melodic events in speech.

Previous research produced interesting, even if sometimes contradictory data from attempts to compare various languages. Mennen, Schaeffler and Docherty (2007) observed a 2.2 ST difference between German and English spoken texts measured by 80% range (referred to as PER in the present study). Their sample was relatively small, but methodologically well managed. Later, in a larger study with different methodology the result was not confirmed (Andreeva et al., 2014). Volín, Poesová & Weingartová (2015) also found a 2 ST difference for 80% range, this time in a large and carefully controlled speech sample comparing Czech and English news reading.

The research in foreign-accented speech is naturally attempting to exploit the fact that languages might differ in their typical overall pitch ranges. However, the influence of the speech style, pragmatic context and affective charge of the communicative situations seems to be stronger than the global inherent tendency in the language. That poses specific demands on the research design. Not only the spoken texts, but also the recording conditions should be made comparable. Even the personality of the experimenter who is collecting the data is perhaps not to be underestimated. The present study tried to embrace these requirements and found that Anglophone speakers who learn Czech as a foreign language tend to use pitch ranges narrower than the ranges used in their mother tongue and often even narrower than those used in their target Czech.

Since the narrower pitch ranges have been reported in foreign-accented speech elsewhere, we might speculate that rather than a mutual interference of two intonational phonologies there is a unique trend in operation: the uncertainty of an L2 learner leads to pitch range compression. In a way, this could be a universal feature of foreignness – at least for situations in which the L2 learners “struggle”. As Major quite convincingly demonstrated, listeners unfamiliar with a language evaluate the degree of accentedness in it similarly to people who know the language (Major, 2007). Maybe pitch range is one of the cues. We should be cautious about the term *uncertainty* used above, though. Most probably it is a complex affective and cognitive structure that will be difficult to define and measure. (This might be further complicated by the fact that the subjects often unconsciously deny the phenomenon’s existence, or the opposite – they consciously claim it while in reality they do not possess it.)

As to further methodological problems, current practice in pitch range quantification seems to be quite crude. The suggestions of Hirst (2003, 2013) are not widely accepted and, furthermore, they do not reflect syntactic/semantic architecture of utterances. A more linguistically motivated approach would require investigation of the local pitch ranges. One such possibility was suggested by Patterson and Ladd (1999), but, to the best of our knowledge, not pursued any further, perhaps because of its labour-extensive nature. Yet it is known that apart from setting the global span, speakers also expand or compress their pitch ranges within utterances depending on the actual communicative situation. This invites quantitative research and promises interesting results if adequate methods are found.

ACKNOWLEDGEMENTS

This study was supported by the Internal Grant Scheme at Metropolitan University Prague under subsection no. VVZ-52-01. The authors would also like to express their sincere thanks to all the volunteers who offered their voices for the purpose of intonation research.

REFERENCES

Abercrombie, D. (1956). *Problems and Principles. Studies in the Teaching of English as a Second Language*. London: Longmans, Green.

Anderson-Hsieh, J., Johnson, R. & Koehler, K. (1992). The relationship between native speaker judgements of nonnative pronunciation and deviance in segmentals, prosody and syllable structure. *Language Learning*, 42, 529–555.

Andreeva, B., Demenko, G., Wolska, M., Möbius, B., Zimmerer, F., Jügler, J., Oleskovicz-Popiel, M. & Trouvain, J. (2014). Comparison of pitch range and pitch variation in Slavic and Germanic Languages. In: *Proceedings of Speech Prosody 2014*, 776–780.

Boersma, P. & Weenink, D. (2014). Praat: doing phonetics by computer. Version 5.4.06, retrieved from http://www.praat.org/.

Derwing, T., Munro, M. J. & Wiebe, G. (1998). Evidence in favor of a broad framework for pronunciation instruction. *Language Learning*, 48, 393–410.

Derwing, T. M. & Rossiter, M. J. (2003). The Effects of Pronunciation Instruction on the Accuracy, Fluency, and Complexity of L2 Accented Speech. *Applied Language Learning*, 13, 1–17.

Field, J. (2005). Intelligibility and the listener: the role of lexical stress. *TESOL Quarterly*, 39 (3), 399–423.

Gilbert, J. (2014). Myth 4: Intonation is hard to teach. In: L. Grant (Ed.), *Pronunciation Myths: Applying Second Language Research to Classroom Teaching*, 107–136. Ann Arbor, MI: University of Michigan Press.

Grossberg, S. (2003). Resonant neural dynamics of speech perception. *Journal of Phonetics*, 31, 423–445.

Hirst, D. J. (2003). Pitch parameters for prosodic typology: A preliminary comparison of English and French. In: *Proceedings of 15th ICPhS*, 1277–1280.

Hirst, D. J. (2013). Melody metrics for prosodic typology: comparing English, French and Chinese. In: *Proceedings of Interspeech 2013*.

Hirst, D. J. & Di Cristo, A. (1998). *Intonation Systems*. Cambridge: Cambridge University Press.

Jilka, M. (2000). *Testing the contribution of prosody to the perception of foreign accent.* Dissertationsschrift zur Dr. phil, Fakultät für Philosophie der Universtität Stuttgart.

Jun, S. A. & Oh, M. (2000). Acquisition of second language intonation. In: *Proceedings of ICSLP (Interspeech)*, 73–76.

Kang, O., Rubin, D. & Pickering, L. (2010). Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English. *The Modern Language Journal*, 94(4), 554–566.

Keating, P. & Kuo, G. (2012). Comparison of speaking fundamental frequency in English and Mandarin. *Journal of Acoustical Society of America*, 132, 1050–1060.

Ladd, D. R. (2008). *Intonational Phonology*. Cambridge: Cambridge University Press.

Lee, S. (2014). The realization of French rising intonation by speakers of American English. In: *Proceedings of Speech Prosody 2014*, 762–765.

Major, R. C. (2007). Identifying a foreign accent in an unfamiliar language. *Studies in Second Language Acquisition*, 29, 539–556.

Mennen, I., Schaeffler, F. & Docherty, G. (2007). Pitching it differently: a comparison of the pitch ranges of German and English speakers. In: *Proceedings of the XVIth ICPhS*, 1769–1772.

Patterson, D. (2000). *A Linguistic Approach to Pitch Range Modelling*. unpublished doctoral dissertation, University of Edinburgh.

Patterson, D. & Ladd, D. R. (1999). Pitch range modeling: Linguistic dimensions of variation. In: *Proceedings of the XIVth ICPhS*, 1169–1172.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40, 227–256.

Van Engen, K. J. & Peelle, J. E. (2014). Listening effort and accented speech. *Frontiers In Human Neuroscience*, 8, 1–4.

Volín, J. & Bartůňková, H. (2015). Assets and liabilities of simple descriptors of fundamental frequency tracks. In: Niebuhr, O. and Skarnitzl, R. (Eds.), *Tackling the Complexity in Speech*, 147–161. Praha: Faculty of Arts Press.

Volín, J., Poesová, K. & Weingartová, L. (2015). Speech melody properties in English, Czech and Czech English: Reference and interference. *Research in Language*, 13(1), 107–123.

RESUMÉ

Studie se zabývá intonačním rozpětím v mluvě anglofonních cizinců, kteří se učí česky. Osm mužů a osm žen, jejichž mateřským jazykem je angličtina, ale kteří jsou schopni používat češtinu alespoň na úrovni B1 ESR, bylo požádáno, aby se seznámili s textem rozhlasového zpravodajství a přečetli jej na mikrofon v nahrávacím studiu Fonetického ústavu FF UK. V nahrávkách byly identifikovány nádechové úseky a čtyři typy korelátů intonačního rozpětí byly změřeny v křivkách průběhů základní hlasové frekvence F0. Tyto hodnoty byly porovnány s referenčními hodnotami pro češtinu a angličtinu známými z literatury. Ukázalo se, že Angličané a Američané ve své češtině používají užší intonační rozpětí než v angličtině a u některých ukazatelů dokonce i užší než bylo naměřeno pro češtinu. Ve shodě s výsledky podobných studií zahraničních je možno se domnívat, že návyk na určité intonační rozpětí se nepřenáší do osvojovaného cizího jazyka. Intonační rozpětí spíše odráží určitou živost sdělení nebo jistotu, s jakou mluvčí daný jazyk ovládá. I když se tedy české mluvené texty vyznačují užším intonačním rozpětím než odpovídající texty anglické, hodnoty pro češtinu s anglickým přízvukem se nenacházejí mezi hodnotami pro oba jazyky.

Jan Volín
Institute of Phonetics
Faculty of Arts, Charles University
jan.volin@ff.cuni.cz

Damien Galeone, Wesley Johnson
Metropolitan University Prague

2017 ACTA UNIVERSITATIS CAROLINAE PAG. 65–78
PHILOLOGICA 3 / PHONETICA PRAGENSIA

REPEATS IN ADVANCED SPOKEN ENGLISH OF LEARNERS WITH CZECH AS L1

TOMÁŠ GRÁF

ABSTRACT

The article reports on the findings of an empirical study of the use of repeats – as one of the markers of disfluency – in advanced learner English and contributes to the study of L2 fluency. An analysis of 13 hours of recordings of interviews with 50 advanced learners of English with Czech as L1 revealed 1,905 instances of repeats which mainly (78%) consisted of one-word repeats occurring at the beginning of clauses and constituents. Two-word repeats were less frequent (19%) but appeared in the same positions within the utterances. Longer repeats are much rarer (<2.5%). A comparison with available analyses show that Czech advanced learners of English use repeats in a similar way as advanced learners of English with a different L1 and also as native speakers. If repeats are accepted as fluencemes, i.e. components contributing to fluency, it would appear clear that many advanced learners either successfully adopt this native-like strategy either as a result of exposure to native speech or as transfer from their L1s. Whilst a question remains whether such fluency enhancing strategies ought to become part of L2 instruction, it is argued that spoken learner corpora also ought to include samples of the learners' L1 production.

Key words: fluency, disfluency, repeats, repetitions, L2 fluency, fluencemes

1. Introduction

The continuous flow of spontaneous speech production is frequently patterned with performance phenomena which include especially lexical and non-lexical fillers, pauses, drawls, truncations, false starts, self-corrections, editing expressions and repeats. These phenomena are understood to relieve the pressure of online planning as their production helps the speaker to acquire time for planning in order to align mental planning with the physical aspects of speech production, or as is frequently the case for L2 speakers for choosing the appropriate form for the message which is being relayed. Unless these elements are too audible (e.g. loud filled pauses) or unusually frequent, they often go unnoticed and do not disturb the listener, who might, in fact, be informed by their pres-

https://doi.org/10.14712/24646830.2017.34

ence that the speaker is aiming to carry on talking or that he is in the process of finding the desired content or form. However, as these features are semantically superfluous to the overall utterance they are generally labelled as disfluencies or dysfluencies and thus carry a rather negative connotation, being seen as elements which disrupt speech fluency.

In a seminal study of hesitation phenomena, Maclay and Osgood (1959) refer to disfluencies as "hesitation errors" and to those who produce fewer of them as "better speakers" (p. 35). As research develops, disfluencies come to be seen not only less negatively but also as essential and natural components of speech production. Fox Tree and Clark (1997) assume that they present a strategy whose function is to solve processing difficulties (see also Clark, 2002). Such a view is justifiable if we consider how ubiquitous disfluencies are (Biber et al., 1999; Kjellmer, 2008). Clark and Wasow (1998) link disfluencies with planning problems and explore them as evidence of planning. This view is developed by Segalowitz (2010) who analyses Levelt's speech production model[1] and identifies within it seven "vulnerability points for fluency" (p. 9). These are defined as "critical points where underlying processing difficulties could be associated with L2 speech dysfluencies". Segalowitz's approach presents a deep-structure model for disfluencies in that he does not provide analyses of concrete instances of disfluencies but instead focuses on identifying where in the model problems may occur. As it is based on Levelt, it is not language-specific (we do not know how disfluencies are realised in different languages) and does not offer a surface-structure view which would investigate concrete realisations of these problems in terms of their qualitative, temporal, or locational characteristics. The nature of the problems is primarily in encoding on grammatical, lexical or phonological level. Segalowitz thus treats disfluencies as predominantly hesitational in nature, but he also acknowledges the existence of vulnerability points in the conceptual preparation phase and in the self-perception processes (see also Li and Tilsen's (2015) discussion of whether disfluencies stem from planning or monitoring).

Skehan (2003) offers a more surface-structure view in his tri-partite model of fluency which sees speech fluency as a sum of speed fluency, breakdown fluency and repair fluency. In this model, some disfluencies are of hesitational nature, while others, such as repeats, aim to repair what has broken down in order to restore the impression of continuous speech. They are used as a communicative strategy, and as is suggested by Rühlemann (2006) should not be called dysfluencies – as dys- implies abnormality – but rather disfluencies, using a weaker reversative prefix. Götz (2013) goes even further in introducing the concept of a fluenceme, which is any component of speech which contributes to either productive or perceptive fluency. In her model of fluency, those features traditionally labelled as dysfluent or disfluent are categorized alongside such phenomena as speech rate or n-grams and rather than hesitational are seen as strategic. Such a view, however, fails to acknowledge that not all disfluent behaviour is necessarily strategic.

Whilst disfluencies have a role in helping the speaker to formulate his message, they also have an effect on the recipient and on the process of comprehension. MacGregor et al. (2009) show that in this respect not all disfluencies are the same: filled pauses are processed with greater ease while repeats are more disruptive as structural and semantic

[1] Levelt, W. J. M. (1999). Language Production: A Blueprint of the Speaker. In: Brown, C. & Hagoort, P. (Eds.), *Neurocognition of Language*, 83–122. Oxford: Oxford University Press.

interpretation must be restarted. This may be especially true of non-native-speaker L2 processing as was shown by Voss (1979) who found that hesitation phenomena were sources of perceptual errors and problems for non-native speakers.

Despite their ubiquitous nature, the production of disfluencies may vary from speaker to speaker. This may give rise to different patterns of disfluent behaviour with differences in the type of disfluency used, its frequency or different combinations of disfluent elements. Disfluent behaviour is thus, to a certain extent, seen as speaker-specific. This was observed already by Maclay and Osgood (1959), and more recently for example by Götz (2013), Braun and Rosin (2015) and McDougall et al. (2015). The characteristics of disfluent behaviour have also been shown to be affected by non-linguistic factors such as gender and age (Bortfeld et al., 2001; Longauerová, 2016) or the type of context and the related level of anxiety or stress (Buchanan et al., 2014).

As regards the location of disfluencies within utterances, they frequently occur before long or complex constituents (Kjellmer, 2008; Watanabe et al., 2008), before grammatically complex constituents (Clark and Wasow, 1998), or before low-frequency words (Corley et al., 2007). Arnold et al. (2003) observe that they frequently precede items which are newly introduced into discourse. Biber et al. (1999) note that the location of the different types of disfluencies varies: unfilled pauses tend to separate major syntactic units, filled pauses lesser syntactic units and repeats may introduce any sentential constituent (e.g. a prepositional phrase). They also acknowledge that the location of disfluencies may be affected by cognitive problems resulting from the nature of the task and that cognitively demanding tasks may result in a higher variability in the type, frequency and location of hesitations. An interesting but not an entirely attested hypothesis is that hesitation phenomena are periodically distributed in spoken language production (Merlo & Barbosa, 2010).

The present study investigates the phenomenon of repeats, i.e. segments of speech which are involuntarily repeated in close proximity without adding any propositional content to the message. Along with filled pauses, repeats are amongst the most frequently occurring types of disfluency (Biber et al., 1999), which however need to be distinguished from repetitions, i.e. deliberate repetitions of words or phrases for rhetorical or other reasons. Example (1) is an illustration of a repeat, and whilst example (2) may be used as an illustration of a repetition of the intensifier *very* for added emphasis, it also shows that distinguishing between these two phenomena may be problematic: without access to the recording to judge the intonation we might not be able to determine whether the repetition of the word *very* is for reasons of emphasis or as a result of hesitation or planning difficulties.

(1) *I mean the the play is really great*
(2) *but the language really was very very nice*

In a seminal study, Clark and Wasow (1998) present repeats as analysable units composed of four subprocesses (initial commitment, suspension, hiatus, and restart). The speaker initially commits to a particular constituent, then suspends speech (for reasons of planning or other), he may fill the hiatus phase with a pause (filled or unfilled), and then resumes production by repeating from the start of the constituent. They observe that the most frequently repeated words are those which are at the left-most side of the constituent: in English these positions are frequently occupied with function words rather than lexical ones. This is in line with both earlier (e.g. Maclay & Osgood, 1959) and later findings (Biber et al., 1999; Kjellmer, 2008) which show that the most typically repeated units are pronouns, articles, prepositions and contracted forms. Clark and Wasow (ibid.) claim that speakers produce repeats because they prefer to deliver continuous speech and therefore after the suspension of speech they start anew. Whilst this is a plausible hypothesis, it fails to explain why repeats are not produced by all speakers and after all points of speech suspension.

Within the context of non-native speech production, Lennon (1990) and Freed (2000) studied various aspects of fluency on a small sample of speakers in a study-abroad context. Whilst they do not provide a detailed analysis and typology of repeats, they observe changes in the frequency of disfluencies including repeats following the speakers' stay in an English-speaking country. Contrary to expectation, this change does not necessarily mean a decrease, which leads Freed to speculate whether the higher frequency of repeats may not be linked to the growing sophistication of the speakers' speech as a result of study abroad.

To date, the most thorough analysis of repeats used by non-native speakers is offered by Götz (2007, 2013). She compares German advanced learners of English with British native speakers and establishes patterns of overuse and underuse of different types of repeats based on Biber's et al. (1999) typology. These results are, however, hard to interpret as the studies do not describe in detail the methodological aspects of locating and classifying the repeats she was working with (the same is true of the above-mentioned studies by Lennon and Freed).

The current study aims to explore quantitative and qualitative aspects of the use of repeats by Czech advanced speakers of English and contribute to the ongoing discussion of the nature of disfluencies in non-native speaker spontaneous speech production. We are specifically interested in whether Czech advanced learners of English show any similarities in their use of repeats to those described in literature on native and non-native use of these disfluencies. This study thus extends what we view as a relatively underresearched area of L2 fluency and disfluency research.

2. Method

The data for the current study derives from the Czech subcorpus of the Louvain International Database of Spoken English Interlanguage (henceforth LINDSEI_CZ) (Gráf, 2017) which contains 50 approximately 15-minute recordings of advanced[2] English learners with Czech as their L1. This amounts to almost 13 hours of recorded material. The learners form a relatively homogeneous group of speakers of similar age (they were all 3rd- or 4th-year university students of English and American Studies), with 43 female and 7 male speakers. Such a homogeneous group does not allow for the exploration of age or gender related effects on fluency as were mentioned above. The orthographic

[2] LINDSEI uses institutional definition of advancedness (Gráf, 2015) and consequently the actual level of proficiency may not be the same for all of its speakers.

transcriptions of the recordings include disfluencies (filled and unfilled pauses, repeats, truncations and drawls) which are counted as words. LINDSEI_CZ contains 123,761 words, of which 95,904 are words produced by the learners. The remaining 27,857 words uttered by the interviewer have not been included in the analysis.

To tag the instances of disfluencies I developed a simple interlinear, incremental tagging system (see Table 1 for examples). The first position of each tag contains the identification of the disfluency type (R = repeat, FS = false start, SC = self-correction). The second position is numerical and describes the length of the repeated phrase. Number 1 thus denotes a repetition of one word, number 2 of two words etc. The third position is numerical and expresses the number of times the phrase is repeated. The fourth position uses letters to encode the part of speech and various subtypes[3]. The fourth position is primarily used with repeats involving one word only. The fifth position is optional and helps distinguish subtypes (e.g. repetitions for rhetorical or discourse purposes).

Table 1. Examples of tags for repeats and their decoding

Example of a tag	Meaning of the tag
<R_1_3_P> I I I wouldn't do it myself	R = repeat, 1 = repeating one word, 3 = occurring 3 times, P = pronoun
which is . similar . (eh) <R_2_2> in many in many ways	R = repeat, 2 = repeating two words, 2 = occurring twice

In order to increase the reliability of the identification process I compiled a computer script for the automatic retrieval and tagging of repeated sequences. The script ignored any intervening pauses and fillers (and their combinations) so that sequences such as *I (erm) I* or *I . I* would still be identified as repeats. This follows Clark and Wasow's conception of repeats as analysable units, and more specifically the notion of hiatus, i.e. the space between the suspension and resumption of speech which may be left unfilled but may also be filled with different types of pauses.

Once all repeats were automatically tagged by the script, I listened to the individual files whilst following the tagged transcriptions to check whether the tagging was done correctly. This helped to distinguish between repeats and repetitions (usually disambiguated by intonation), and it also revealed instances in which the occurrence of two identical words next to each other were not cases of repeats. They were cases in which the co-occurring words were not part of the same constituent, as shown in examples (3–5), or sentence (the transcription does not use punctuation).

(3) *the Film Society have got it on on a Friday*
(4) *we went to see it it was Sunday morning*
(5) *we have had compliments from outside companies companies that normally deal with proper commercial cinemas*

[3] Meaning of the codes in the fourth position of the tags: Ad – definite article; Ai – indefinite article; Ao – other determiner; B – preposition; C – conjunction; D – discourse marker; E – existential there; F – filler; G – adverb; Ip – infinitive particle; J – adjective; N – noun; O – other ; P – pronoun; R – rhetorical; V – verb; W – wh-word; X – contraction

Only fully retraced elements were tagged as repeats, thus if the element involved any kind of rephrasing, it was tagged as a false start (FS), as shown in example (6). Also tagged as false starts were all instances in which only a part of the word was repeated as shown in example (7), and in example (8) in which each repetition of the initial syllable is tagged separately as a false start.

(6) *<FS_2> she didn't (eh) she . couldn't agree because*
(7) *she was dissatisfied <FS_1> wi= with the painting*
(8) *I mean <FS_1> av= <FS_1> av= . avoiding conflicts*

As in similar studies (e.g. Maclay & Osgood, 1959; Clark & Wasow, 1998; Götz, 2013) graphic words are counted, which means that contractions are counted as one word (e.g. *I'm*, *it's* etc.). Biber et al. (1999) point out that this procedure is fully justifiable as contractions are processed as single items.

Once the tagging was completed, the files were analysed using AntConc (Anthony, 2014). Excluded from the count were all instances of repetitions for rhetorical (see example (2) above) or discourse purposes, as in example (9), and repetitions of filled pauses.

(9) Interviewer: *are you writing your bachelor's thesis at the moment*
Interviewee: *not yet not yet I . plan to write it . during the[i:] Erasmus so*

3. Results

A total of 2,311 sequences of repeated elements were identified. Once all instances of non-repeats as described in the preceding section were removed, 1,905 repeats remained for our analysis. As is shown in Table 2, more than three quarters (78.27%) of the bulk are formed by one-word repeats. Multi-word repeats are less common, with two-word repeats adding up to 19.3%, three-word repeats to 2.4% and longer repeats to approximately 0.1%.

Table 2. Frequencies of repeats of different length

Length of repeated segment	N	%
One word	1,498	78.3 %
Two words	369	19.3 %
Three words	45	2.4 %
Four words	2	0.1 %
More than four words	0	0 %
Total	1,914	100 %

Following Clark and Wasow's (1998) method, our discussion of repeats does not subcategorize repeats with different types of hiatus or other variations. These are, however, relatively frequent: 20% of instances of one-word repeats include an unfilled pause (as in ex. 10), 5% include a filled pause (ex. 11), 3% include lengthening (ex. 12) and 3% include an extension of a personal pronoun by a contracted form of a copular or auxiliary verb (ex. 13). The situation is similar for multi-word repeats.

(10) *you can really enjoy <R_1_2_Ad> the . the view every morning*
(11) *I'm a huge fan <R_1_2_B> of (erm) . of television series*
(12) <R_1_2_Ad> the: the lady seems to be pleased
(13) *I mean <R_1_2_P> I I've been doing that*

3.1 One-word single repeats

Table 3 shows a breakdown of the <R_1_2> type, when the speaker repeats a single word once. As is pointed by Biber et al. (1999: 1055) this is the most common type of repeat. In the present corpus, 1,349 such instances have been observed. The most commonly repeated elements are pronouns, conjunctions, prepositions, definite articles and contracted forms. These parts of speech also show a very high frequency across the board, as 98% of all of the speakers produced at least one instance of pronoun repetition, 90% of speakers repeat prepositions, 68% conjunctions, 68% contractions, and 66% repeat definite articles.

Table 3. Frequencies of one-word single repeats (tagged as <R_1_2_*>)

Repeated element	Count	%	Speakers involved
Pronoun	470	34.8 %	49 (98 %)
Preposition	164	12.2 %	45 (90 %)
Conjunction	163	12.1 %	34 (68 %)
Definite article	117	8.7 %	33 (66 %)
Contracted form	103	7.6 %	34 (68 %)
Adverb	61	4.5 %	28 (56 %)
Other types	59	4.4 %	32 (64 %)
Verb	53	3.9 %	23 (46 %)
Infinitive particle	42	3.1 %	24 (48 %)
Wh-word	39	2.9 %	24 (48 %)
Adjective	33	2.4 %	20 (40 %)
Noun	13	1.0 %	7 (14 %)
Indefinite article	12	0.9 %	10 (20 %)
Existential there	11	0.8 %	11 (22 %)
Other determiner	9	0.7 %	7 (14 %)
Total	1,349	100 %	

3.2 One-word multiple repeats

Multiple repeats of one word are considerably less common. The corpus contains 140 instances of triple repeats and 8 instances of quadruple repeats. As is shown in Table 4, these are again most frequently repeats of pronouns (40.5%), definite articles (8.8%), conjunctions (8.1%) and prepositions (6.7%), but they occur in a much smaller selection of speakers: except pronouns which were repeated by 58% of speakers, all of the other types are repeated by fewer than 20% speakers.

Table 4. Frequencies of one-word multiple repeats (tagged as <R_1_3/4_*>)

Repeated element	Count	%	Speakers involved
Pronoun	60	40.3 %	29 (58 %)
Definite article	12	8.1 %	9 (18 %)
Conjunction	12	8.1 %	6 (12 %)
Preposition	10	6.7 %	9 (18 %)
Other types	12	8.1 %	8 (16 %)
Wh-words	8	5.4 %	5 (10 %)
Infinitive particle	6	4.0 %	4 (8 %)
Adjective	6	4.0 %	4 (8 %)
Contraction	6	4.0 %	6 (12 %)
Verb	4	2.7 %	4 (8 %)
Adverb	2	1.3 %	3 (6 %)
Indefinite article	2	1.3 %	1 (2 %)
Noun	1	0.7 %	1 (2 %)
Pronoun (repeated 4 times)	8	5.4 %	6 (12 %)
Total	149	100.00 %	

3.3 Multi-word repeats

Multi-word repeats detected in our corpus include 370 instances of two-word repeats and 45 instances of three-word repeats. The majority (241 instances, 65.14%) of our two-word repeats involve a subject followed by different types of complementation. As we can see in Table 5, the most frequent types are subject + copular verb (40.2%, see ex. 14), subject + auxiliary/modal verb (24.1%, see ex. 15), subject + lexical verb (12.9%, see ex. 16), and a combination of subject preceded by another word (17.43%), such as a conjunction as in ex. (17). Other instances are marginal.

(14) *<R_2_2> it was it was just the inability to act*
(15) *<R_2_2> I am I am planning my next visit to*
(16) *<R_2_2> we see . we see children from the whole world*
(17) *<R_2_2> when she when she actually sees the painting*

Two-word repeats frequently involve a verb (218 cases). These mostly (88.5%) include the combinations of subject + verb discussed above (see Table 5). Other combinations are rarer, such as verb + preposition (2.75%), copular/auxiliary/modal verb + lexical verb (1.4%), verb + object (2.3%), and to + infinitive (1.4%).

Table 5. Proportional distribution of repeats involving a subject and different types of complementation

	Copular	Auxiliary	Lexical	* + S	Existential	Adverb	Total
Subject	97	58	31	43	7	5	241
%	40.2 %	24.1 %	12.9 %	17.8 %	2.9 %	2.1 %	100 %

Table 6 displays the number of two-word repeats involving a preposition. The most frequently occurring repetitions of this type include the prepositions *in, on, to, of, with* and *as* (examples 18 and 19).

(18) *there are a lot of catchy phrases (erm) <R_2_2> in the in the play*
(19) *but I got to go <R_2_2> on a on a cruiseship*

Table 6. Proportional distribution of repeats involving prepositions + complement

in	on	to	of	with	as	for	from
14	8	7	7	6	4	4	4
23.0 %	13.1 %	11.5 %	11.5 %	9.8 %	6.6 %	6.6 %	6.6 %
at	like	about	through	by	instead of		Total
2	1	1	1	1	1		61
3.3 %	1.6 %	1.6 %	1.6 %	1.6 %	1.6 %		100 %

Table 7 provides an overview of two-word repeats involving a conjunction. The most frequently occurring repetitions here include *wh*-words used as conjunctions (ex. 20), and then the conjunctions *and*, *so, that, as* and *if.* The majority of these repetitions (79%) are the combination of a conjunction followed by a subject (ex. 21), the remaining 21% are used within a nominal phrase (ex. 22).

(20) *it's massive and <R_2_2> when you when you really enter into it*
(21) *she's wearing (er) . a pretty dress <R_2_2> and he and he starts painting*
(22) *between the teacher <R_2_2> and the and the students*

Table 7. Proportional distribution of repeats involving conjunctions

	wh- word	and	so	that	as	if	but	after	before	than	Total
Conj. + wh	25	10	5	4	4	4	2	2	1	1	58
%	43.1%	17.2 %	8.6 %	6.9 %	6.9 %	6.9 %	3.4 %	3.4 %	1.7 %	1.7 %	100 %

With 45 instances, three-word repeats are considerably less frequent. Almost two thirds of them (64.4%) involve a subject and a verb (exs. 23–25), 9% are within a prepositional phrase (ex. 26), and then there are only singular instances of different types (e.g. ex. 27).

(23) *<R_3_2> it would be it would be very easy*
(24) *<R_3_2> it was an it was an awesome experience*
(25) *<R_3_2> I can see I can see the point*
(26) *<R_3_2> in the last in the last picture we can see*
(27) *<R_3_2> what would he what would he do*

3.4 Repeat rates

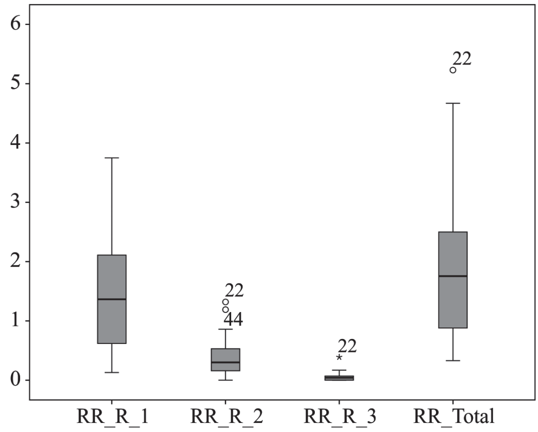
In the following section, I will inspect the frequency of occurrence of repeats as they were produced by the learners. Here a normalized frequency per hundred words (phw) is used, which I will henceforth refer to as "repeat rate". Table 8 shows that the overall repeat rate in the whole corpus is 1.91 repeats phw (SD=1.18), which means that repeats occur once in every 52 words. One-word repeats occur at a rate of 1.47 phw (SD=0.94) (one instance every 68 words), two-word repeats at a rate of 0.37 repeats phw (SD=0.29) (once in every 270 words), and three-word repeats at 0.05 (SD=0.09) (once in every 2,000 words).

Table 8. Mean repeat rates per hundred words (phw)

Type of repeat	Repeat rate phw	SD
One-word	1.47	0.94
Two-word	0.37	0.29
Three-word	0.05	0.09
Total	1.91	1.18

The large standard deviations, however, indicate that there is a large inter-speaker variability in the production of repeats. Whilst the least disfluent speaker repeats at a rate of 0.33 repeats phw (one repeat every 303 words), the most disfluent one repeats at a rate of 5.23 repeats phw, producing one repeat every 19 words. Figure 1 provides a comparison of the repeat rates for the different types of repeats.

Figure 1. Ranges and the distribution of repeat rate values for one-word repeats (RR_R_1), two-word repeats (RR_R_2), three-word repeats (RR_R_3), and all repeats (RR_Total).



The values range from 0.32 to 5.12 repeats phw for one-word repeats, from 0 to 1.38 for two-word repeats, and from 0 to 0.39 for three-word repeats. All of the 50 speakers produced at least one instance of one-word repeat, 47 speakers produced two-word repeats, and 30 speakers produced three-word repeats.

4. Discussion

The purpose of this study was to investigate the use of repeats in the spontaneous spoken production of advanced learners of English. In particular, I investigated the types of repeated words and established a typology of repeats which occur in their spoken production. The most frequently occurring types of repeats in our corpus are one-word repeats. Within this group the most frequently repeated words are pronouns (including those with a contracted form with a verb), and especially personal pronouns. Taking into account the typical structure of English sentences, this finding confirms those of many previous studies which indicate that the use of repeats is a strategy connected with the planning of an utterance where most of the planning pressure is at the beginning of an utterance. Other major types of repeats contain articles or prepositions, which implies that planning pressure also increases and time needs to be gained at the beginning of noun or prepositional phrases. The last major category includes the use of conjunctions at the beginning of a clause, and also – if less frequently – within a noun phrase.

These results are in line with those of Biber et al. (1999)[4], who, however, investigated the use of repeats in native-speaker English. It is interesting to observe that based on this comparison our advanced learners use similar strategies and with similar frequency as native speakers. The only comparable study exploring the phenomenon of advanced learner English is one carried out by Götz (2013), who in her own corpus of 50 advanced learners of English with L1 German observed the same types of repeats. The high dispersion in her data leads Götz (p. 109) to consider whether repeats are a fluency enhancing strategy which has been adopted only by the more advanced learners. However, she records a very similar dispersion in a parallel corpus of native speakers, which would rather seem to imply that different speakers might use different strategies to gain time for planning speech. Repeats are only one of these strategies, others including, for example, varying speech rate or the use of different pause, false-start or self-correction patterns. This area requires further investigation.

As regards multi-word repeats, the majority of them involve a subject and can thus again be found most frequently at the beginning of clauses. Our results cannot be compared here with other studies as the three main studies referred to above[5] deal only with one-word repeats.

In agreement with Clark and Wasow (1998), Biber et al. (1999) and Götz (2007), our repeats are also accompanied by other types of disfluencies (in 15% of the cases), especially pauses or syllable lengthening. These occur either in the hiatus (i.e. between the repeated segments) but sometimes also before or after it. This illustrates that repeats themselves are not always sufficient means of gaining planning time and other strategies are adopted by the speakers in combination. Multiple-repetitions, i.e. those that involve more than two-fold repeats are fairly infrequent.

Whilst repeats were found to be used by all of our speakers, the large dispersion in their use shows that the group is rather heterogeneous and the strategy is not used by all speakers to the same degree. Further investigation is warranted especially with regard to

[4] Biber et al. (1999) unfortunately do not report on the dispersion of their data.

[5] i.e. Maclay and Osgood (1959), Biber et al. (1999), and Götz (2013)

their use of alternative choices of speech planning strategies, and this also raises a question whether the use of repeats is an area of pedagogical implications, and – more specifically – whether learners ought to be taught how to use repeats and fluency enhancing strategies in general.

5. Conclusion

This study has shown that as much as 2.5% of spontaneously produced speech in L2 learners is accounted for by the repetition of segments. This repetition might be seen either as a type of disfluency or as a fluency-enhancing strategy which allows the speaker to gain time for planning speech. The typology of these repeats has revealed that repeats are predominantly used at the beginnings of clauses or of nominal/prepositional phrases, where planning pressure is felt most acutely, and that the learners thus feel the need to plan not only at the beginning of clauses but also at the beginning of other constituents. More research is needed to explore the differences in the location of repeats produced by learners and native speakers.

However, not all of the learners appear to make use of this strategy and future studies of this matter should concentrate on finding which strategies are used as alternatives. Also, correlations can be sought between the use of repeats and proficiency, trying to determine whether more advanced learners use fluency enhancing strategies more effectively. Further research ought to be carried out investigating and explaining the similarities between the use of repeats in native and learner language. It would also seem worth our attention to see whether the use of repeats by L2 speakers mirrors their use of this strategy in their L1, and whether, indeed, this might be a specific area of language transfer. To this purpose, it would appear beneficial if learner corpora contained also samples of the participants' L1.

Previous studies of repeats in native speech show these to be a natural component of everyday speech. The present study shows that they are also frequent in L2 advanced speech. It is likely that the use of such time-gaining strategies positively affects fluency, and an important question must thus be raised whether L2 learners ought to consciously adopt such strategies and whether they can be helped in this process by explicit instruction.

ACKNOWLEDGEMENT

This research was supported by the Charles University project Progres 4, *Language in the shiftings of time, space, and culture*.

REFERENCES

Anthony, L. (2014). *AntConc (Version 3.4.3)*. Tokyo: Waseda University. Retrieved from http://www.laurenceanthony.net/

Arnold, J. E., Fagnanon, M. & Tanenhaus, M. K. (2003). Disfluencies Signal Theee, Um, New Information. *Journal of Psycholinguistic Research*, *32*(1), 25–36.

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Harlow, England; New York: Pearson Education ESL.

Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F. & Brenan, S. E. (2001). Disfluency Rates in Conversation: Effects of Age, Relationship, Topic, Role, and Gender. *Language and Speech*, *44*(2), 123–147.

Braun, A. & Rosin, A. (2015). On the Speaker-Specificity of Hesitation Markers. In: *Proceedings of the 18th International Congress of Phonetic Sciences. (ICPhS 2015)*. Glasgow: University of Glasgow.

Buchanan, T., Laures-Gore, J. S. & Duff, M. C. (2014). Acute stress reduces speech fluency. *Biological Psychology*, *97*, 60–66.

Clark, H. H. (2002). Speaking in time. *Speech Communication*, *36*(1), 5–13.

Clark, H. H. & Wasow, T. (1998). Repeating words in spontaneous speech. *Cognitive Psychology*, *37*(3), 201–242.

Corley, M., MacGregor, L. J. & Donaldson, D. I. (2007). It's the way you, er, say it: Hesitations in speech affect language comprehension. *Cognition*, *105*, 658–668.

Fox Tree, J. E. & Clark, H. H. (1997). Pronouncing 'the' as 'thee' to signal problems in speaking. *Cognition*, *62*(2), 151–167.

Freed, B. F. (2000). Is fluency in the eyes (and ears) of the beholder? In: H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 243–265). Ann Arbor: University of Michigan Press.

Götz, S. (2007). Performanzphänomene in gesprochenem Lernerenglisch: Eine korpusbasierte Pilotstudie. *Zeitschrift Für Fremdschprachenforschung*, *18*(1), 67–84.

Götz, S. (2013). *Fluency in Native and Nonnative English Speech*. Amsterdam; Philadelphia: John Benjamins Publishing Company.

Gráf, T. (2015). Accuracy and fluency in the speech of the advanced learner of English (Unpublished PhD Thesis). Charles University, Prague, Czech Republic.

Gráf, T. (2017). *Korpus LINDSEI_CZ*. Praha: FF UK.

Kjellmer, G. (2008). Self-repetition in spoken English discourse. In: T. Nevalainen & I. Taavitsainen (Eds.), *The dynamics of linguistic variation: corpus evidence on English past and present; [... a selection of articles based on papers presented at the 27th Conference of the International Computer Archive of Modern and Medieval English (ICAME) in Helsinki in May 2006]*. Amsterdam: Benjamins.

Lennon, P. (1990). Investigating Fluency in EFL: A Quantitative Approach*. *Language Learning*, *40*(3), 387–417.

Li, J. & Tilsen, S. (2015). Phonetic evidence for two types of disfluency. In: *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS 2015)*. Glasgow: University of Glasgow.

Longauerová, R. (2016). *Sociophonetic study of dysfluent behaviour in native English speakers* (Unpublished BA thesis). Charles University, Prague.

MacGregor, L. J., Corley, M. & Donaldson, D. I. (2009). Not all disfluencies are are equal: The effects of disfluent repetitions on language comprehension. *Brain and Language*, *111*(1), 36–45.

Maclay, H. & Osgood, C. E. (1959). Hesitation Phenomena in Spontaneous English Speech. *Word*, *15*(1), 19–44.

McDougall, K., Duckworth, M. & Hudson, T. (2015). Individual and Group Variation Dysfluency Features: A Cross-Accent Investigation. In: *Proceedings of the 18th International Congress of Phonetic Sciences. (ICPhS 2015)*. Glasgow: University of Glasgow.

Merlo, S. & Barbosa, P. A. (2010). Hesitation phenomena: a dynamical perspective. *Cognitive Processing*, *11*(3), 251–261.

Rühlemann, C. (2006). Coming to terms with conversational grammar: 'Dislocation' and 'dysfluency'. *International Journal of Corpus Linguistic*, *11*(4), 385–409.

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. New York: Routledge.

Skehan, P. (2003). Task-based instruction. *Language Teaching*, *36*(1), 1–14.

Voss, B. (1979). Hesitation phenomena as sources of perceptual errors for non-native speakers. *Language and Speech*, *22*(2), 129–144.

Watanabe, M., Hirose, K., Den, Y. & Minematsu, N. (2008). Filled pauses as cues to the complexity of the upcoming phrases of native and non-native listeners. *Speech Communication*, *50*, 81–94.

RESUMÉ

Spontánní mluvený projev je charakteristický mj. tím, že se v něm vyskytuje velká řada tzv. disfluencí, např. opakování slov, užívání pauz, falešných začátků či přeformulování částí promluv. Jde o jevy, které mohou být chápány jako prvky narušující plynulý tok řeči, ale zároveň mohou být považovány za strategie, které plynulý tok řeči podporují, a to především tím, že mluvčímu umožňují získat čas na plánování promluvy.

Předložená studie se zabývá výzkumem jednoho z nejčastějších typů disfluencí, a to opakování slov. Na základě mluveného žákovského korpusu pokročilé angličtiny mluvčích s češtinou jako mateřským jazykem zkoumá spontánní projev (o celkové délce téměř 13 hodin) 50 studentů anglistiky na FF UK. V tomto vzorku bylo identifikováno 1905 případů opakování. Téměř 80 % z nich tvoří opakování jednoho slova, téměř 20 % opakování dvouslovných segmentů a pouze asi 2 % opakování segmentů delších.

Kvalitativní analýzou bylo zjištěno, že opakování nejčastěji obsahují osobní zájmena, spojky a předložky, což vypovídá o tom, že mluvčí této strategie využívají na začátku vět a větných frází, kde je pociťován největší tlak na plánování promluvy. Tyto výsledky jsou srovnatelné se studiemi projevu rodilých mluvčích i mluvčích angličtiny jako cizího jazyka s jiným mateřským jazykem (němčinou).

Pokud vnímáme opakování jako účinnou strategii při řečovém managementu, nabízí se otázka, zdali tyto strategii nemají být součástí výuky cizích jazyků.

Tomáš Gráf
Department of English Language and ELT Methodology
Faculty of Arts, Charles University
tomas.graf@ff.cuni.cz

APPEAL AND DISREPUTE OF THE SO-CALLED GLOBAL RHYTHM METRICS

JAN VOLÍN

ABSTRACT

Since the late 1990's correlates of rhythm classes of languages have been profusely used to search for differences between languages, dialects, speaking styles, degree of foreign accents, etc. Over the years the original attractiveness of the metrics has been replaced with suspicion and, occasionally, even fierce criticism. Among many reservations the critics argue that the metrics are only based on durational measures ignoring other dimensions of prominence, and they are considerably influenced by local temporal variation in utterances. We argue that the metrics could still be exploited in speech research as long as we do not expect them to reflect "speech rhythm" and as long as the proper account of their use is supplied. This study provides simulations to demonstrate the behaviour of the most commonly utilised metrics, and presents representative measurements of some Czech and English speech recordings under several conditions.

Key words: speech rhythm, rhythm classes, rhythm configurations, temporal structure, global metrics

1. Introduction

Rhythm is generally defined as a flow of contrasts in time with perceived regularity. There are countless examples of the importance of such contrast alternations in human lives. Everything we encounter takes place in time and events keep alternating. Thus, the distribution of differing, contrastive events in time is a fundamental, omnipresent attribute of the world as we know it. The question mark hangs over the term *regularity*, and we will return to it below since it seems to underlie most of the dilemmas current research scene faces in connection with the concept of rhythm.

The rhythm of speech remained excommunicated from linguistic research for a very long time. It was seen as pure means of ornamentation, and since there was no clear effect it would have on intellectual meanings of words with which phonology was pre-occupied, it was left to versologists to study. However, over the time the attitudes have changed dramatically and recent decades have brought international workshops, special issues of journals and dedicated sessions at major scientific conferences concentrating on rhythm.

https://doi.org/10.14712/24646830.2017.35

It is acknowledged that speech with natural rhythm (i.e., typical for the given language and speaking style) is processed more economically by our brains than speech with less common or less predictable rhythmic patterns (e.g., Huggins, 1979; Buxton, 1983; Quené & Port, 2005). Grossberg (2003) provides a feasible neuro-physiological explanation of this effect and useful hints from the same domain are supplied by Ghitza and Greenberg (2009) as well. Hove and Risen (2009) demonstrated the link between rhythm and social cohesion: they showed that shared rhythmic experience increases the mutual positive perception of individuals. A similar study with four-year old children is equally convincing (Kirschner & Tomasello, 2010) in demonstrating the affiliation effects of synchronized rhythmic activities. It comes as no surprise then that the speech styles which are connected with attempts to convince listeners about certain 'truths' are more rhythmical than ordinary conversational speech (Knight & Cross, 2012).

Despite the current generally positive acceptance of the rhythm-related topics, there is also some scepticism or frustration being expressed. After decades of relatively intensive research, there is still no comprehensive model of speech rhythm. Most studies deal with incomplete concepts even if they attempt at framing them into larger theories. Nevertheless, current empirical hypothesis testing is still exciting and inspires advancement in the research area.

An example of how frustration may lead to a denial of speech rhythm and still contribute to the development in the field is the recent article by Nolan and Jeon (2014), who even argue that speech is anti-rhythmic. Their account is not purely provocative, although the desire to stir discussion rather than to present a balanced realistic view might be felt from some of the propositions. The authors covertly equal rhythm with some sort of *neat objective alternations* (overtly only exceptionally, e.g., Nolan & Jeon, 2014: 7), which inevitably leads to its denial in the day-by-day use of speech. It is the problem of regularity advertised in the first paragraph of this section that causes the trouble. I propose that rather than inventing terms for various types of rhythm (contrastive vs. coordinative in Nolan and Jeon's case) it might suffice to recall the relatively old distinction between *meter* and *rhythm* (e.g., Gorow, 2000: 208). If we admit that *rhythm* refers to specific configurations of contrasts, whereas *meter* is an abstract uniform skeleton with which configurations of contrasts might be coupled, then there is no need for a denial of rhythm in speech. Instead, we might argue that there is a very loose meter which does not adhere to the objective physical time.

Even music, whose compulsory feature is regularization of intervals in both frequency and time domains, avoids monotony or repetitiveness in most of its instances. (We tend to praise music which is more varied and to condemn music that is too monotonous). It needs meter to allow for coordination of participating musicians in their joint production. In addition, there is the desire to get the listener entrained. What listeners often do is they mimic the repetition of some of the underlying beats by movements such as claps and foot stamps. This way they create some sort of crude skeleton of what they hear and the outcome is usually closer to the meter than to the surface rhythm.

Speech clearly cannot afford repeating primitive patterns. Monotony or repetitiveness of simple patterns would constrain its readiness to express a variety of meanings, which is its most precious attribute. Research in rhythm of speech should focus on describing the flow of particular configurations in a given language rather than on seeking simple primitive regularity. Separating the concept of meter from rhythm might help to avoid the search for a new term to replace the word rhythm in speech sciences.

The reasoning above also explains why computational techniques that were originally proposed as *rhythm metrics* should not be termed so. First, the early proposals indicated that it is not rhythm, but rhythm classes that correlate with the measures. Second, even *rhythm-class metrics* would not be a satisfying term, since the calculations are mostly based on durational measures only, without any perceptual normalizations. Plain physics cannot substitute for psychoacoustics of durations, let alone for prominence phenomena based on interplays of pitch movements, loudness and timbre variations. Just to mention a few recent examples, Barry, Andreeva and Koreman (2009) demonstrated that pitch changes could influence rhythm judgements to a considerable extent. Cumming designed experiments that highlighted the dependence of the sensitivity to various prominence cues on the native language of the speakers (Cumming, 2011). In her data the French processed tonal and temporal prominence features differently from the German. Brugos and Barnes (2014) cite abundant research in psychoacoustics that exposes the interdependency of pitch and duration percepts, even if in their own carefully prepared experiment the mutual influence was less convincing. These and other objections lead to a relatively wide spectrum of attitudes towards the metrics, which spans from favourable acceptance (e.g., Dankovičová & Dellwo, 2007; White et al., 2007; O'Rourke, 2008; Kinoshita & Sheppard, 2011) through moderate doubts (Loukina et al., 2011; Mariano & Romano, 2011) to categorical refusal (Kohler, 2009; Nolan & Jeon, 2014). A gradual shift from acceptance to refusal over time can be sensed. We suggest that both attitudinal opposites should be reconsidered.

The metrics became quite attractive for several reasons, two of which seem to be most evident. First, they looked exact and sophisticated. They could provide numbers with many positions after the decimal point and as such could help linguistics counter unfair allegations of not being a real 'hard' science. Needless to say that accepting the illusion of exactness is short-sighted since numbers per se are neither accurate nor inaccurate. Second, the metrics seemed to finally corroborate the existence of rhythm classes based on the isochrony of syllables, stress-groups and morae that was seriously doubted in the 1990s. Yet correlations with rhythm classes do not necessarily explain their perceptual foundation. The current debate returns to the rejection of overly simplistic views of mora-, syllable- and stress-timing.

The argumentation above suggests that to talk about the durational metrics as rhythmic measures is apparently misleading. On the other hand, to deny that speech displays non-random alternations of contrasts on several levels is also unhelpful. The potential value of the metrics, then, might be sought in their capacity to capture parameters of temporal organization of speech material. These parameters should ultimately serve to design perceptual experiments that would either confirm or disprove their relevance. However, to make broader use of the potential of the metrics, several conditions must be met. There is a necessity to:

1) replace their misleading label and rather than rhythm metrics refer to them as *durational variation metrics* (DVM),

2) use them on speech material that is thoroughly described with regard to speaking style, context of recording sessions, and articulation rates,
3) avoid using small, inadequate samples of speech material since considerable fluctuations of values have been verified,
4) experiment with the metrics under various conditions to expose their behaviour.

The objectives of the current study match the above stated provisions. Apart from these requirements a few specific goals were set. First, natural speech displays continuous variation which allows for a vast number of durational ratios. These might produce patterns that are difficult to conceptualize. Therefore, the initial analyses will be performed on artificial material (for specifications see below), which renders the effects of various ratios or normalizations clearer. Subsequently, an extensive sample of Czech read monologues will be analysed. These two types of material will be used to provide specific values of DVMs so that various findings in the relatively rich literature can be better compared and appreciated.

Second, the interspeaker and intraspeaker consistency will be captured. Knight (2011) measured the stability of the metrics over time and found out that readings of a text over a period of several days were reasonably consistent. However, she also found that when the text was divided into smaller portions, the consistency over such portions was worryingly low. This warns against putting too much trust in studies that base their claims on five sentences per language.

The data concerning Czech are scarce and unrepresentative although the language displays various interesting features. The Czech vocalic system involves the phonological length of vowels, and consonants are allowed to form clusters. As to the vowels, there are five short and five long monophthongs which, apart from high front vowels, are paired by vowel timbre (Skarnitzl & Volín, 2012). Current Czech also possesses three diphthongs that have durations comparable to long vowels. As to consonant clustering, there may be up to four consonants in syllable onsets and three consonants in syllable codas. However, these extreme clusters are very rare, especially in codas. On the other hand, an onset can meet with a coda of the preceding unit so simple CV alternations are often interspersed with VCCV and VCCCV sequences (Volín & Churaňová, 2010). Reliable reflection of this in terms of durational variation metrics will be provided.

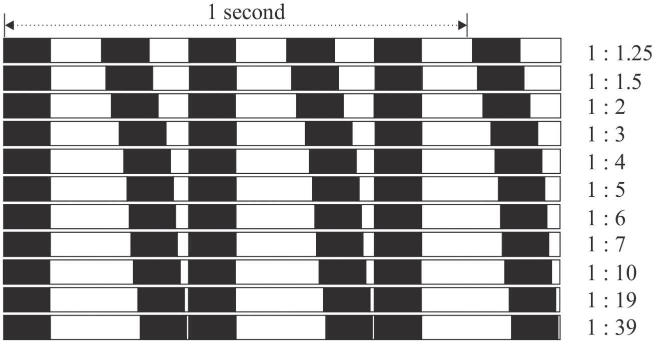
2. Method

2.1 Simulations

The core (or referential) artificial material was set to emulate the articulation rate of 5 syll/s which is a natural and comfortable tempo for most humans (a faster and slower version was also produced). Durations of consonants were kept constant for the sake of clarity, but the behaviour of metrics on vowels would manifest on any units: the metrics are ‘blind’ to what they are supposed to measure. Thus, if the examined units were syllables or stress-groups, as long as the ratios are preserved, the metrics would produce identical results. The intention at this point, however, was to produce material that is based on realistic attributes of speech and, at the same time, is easy to conceptualize.

The material comprised regular alternations of longer and shorter vowels and the manipulated variable was their mutual durational ratio. Figure 1 provides a visualisation of some of the ratios that were used. Apart from the ratios in the diagram, we also used 1:8 and 1:9, and some of the ratios were also used in material simulating slow (3 syll/s) and fast speech (8 syll/s).

Figure 1. Diagram of some of the durational ratios used in the experiment. Black blocks refer to consonants, white blocks to vowels. The length of blocks is proportional to durations.



The *ΔV* and *VarcoV* metrics were in all ratios computed for four different lengths of units. These were chosen to reflect realistic durations of breath-groups in speech, but at the same time to produce integer counts of syllables. They were 1.6 s, 3.2 s, 4.8 s, and 6.4 s.

2.2 Natural material

The natural material entailed recordings of news bulletins from the national broadcaster Czech Radio (Český rozhlas), stations 1 and 2. The speakers were professional news readers (6 women and 6 men) who are generally considered models of standard Czech pronunciation. Their profession requires relatively fast speech rates which, however, are not allowed to interfere with the clarity of pronunciation: the news must be easily intelligible to listeners, who do not see the speaker. The speakers read the bulletins of about 7 paragraphs and 500 words on average (about 2900 vowels and consonants per the bulletin). For most analyses in this study each speaker is represented by one news bulletin, but two of the speakers also provided one extra news bulletin recorded several weeks after the first one. These extra items were used for one of the measurements of intra-speaker consistency.

The recordings were divided into breath-groups, i.e., stretches of speech between the intakes of breath. As the speakers were professionals who prepared their reading beforehand, the breath-group boundaries coincided with major syntactic breaks. Phone boundaries were first estimated (forced aligned) by the Prague Labeller (Volín, Skarnitzl & Pollák, 2005; Pollák, Volín, Skarnitzl, 2007) and then manually corrected for various imprecisions. Altogether, slightly over 41.000 consonantal and vocalic boundaries were checked for further processing. The breath-groups were processed individually, but the arithmetic means calculated later were weighted by their duration. Hence, the shorter units contributed to the overall mean less than longer units.

As the durational variation metrics are dependent on articulation rate, the specification of these must be provided if any cumulative aspect of research is intended. The mean articulation rate in our sample was 6.2 syll/s and 15.25 phone/s. (In line with general convention, articulation rate is calculated after the exclusion of pauses.) The contribution of individual speakers to the mean is displayed in Table 1.

Table 1. Articulation rates of the female (F1 – F6) and male (M1 – M6) speakers in the sample. Values in syllables per second and phones per second are presented.

Speaker	syll/s	phone/s	Speaker	syll/s	phone/s
F1	5.97	14.44	M1	6.59	16.26
F2	5.71	14.20	M2	6.69	16.31
F3	6.20	14.98	M3	6.04	14.58
F4	5.93	14.48	M4	6.23	15.23
F5	6.31	15.80	M5	6.13	15.51
F6	6.44	15.88	M6	6.17	15.27

2.3 Metrics

Seven most commonly used metrics were chosen for the study. They were introduced, for instance, in Low and Grabe (1995), Ramus, Nespor and Mehler (1999), and Dellwo and Wagner (2003), but cf. also Low, Grabe & Nolan (2000), Grabe & Low (2002), or Wagner & Dellwo (2004). The following paragraphs will present their computational bases.

Pairwise variability index (PVI) was originally proposed in its raw version (rPVI) as a mean difference between two successive units:

$$rPVI = \sum_{i=1}^{m=1} \frac{|d_i - d_{i+1}|}{(m-1)}$$

where i is the summation index, m is the number of analysed units in the given breath-group, d_i is the duration of an i-th unit and d_{i+1} is the duration of the subsequent unit. This index is returned in milliseconds and is clearly influenced by the duration of the measured units. Therefore, it will not be used in this study and a normalized version (nPVI) will be exploited instead. The original version (below on the left) was later adjusted to produce a range of values that would be easier to apprehend (below on the right):

$$nPVI = 100 \times \sum_{i=1}^{m=1} \left| \frac{d_i - d_{i+1}}{(d_i + d_{i+1})/2} \right| /(m-1) \quad \rightarrow \quad nPVI = 100 \times \sum_{i=1}^{m=1} \left| \frac{d_i - d_{i+1}}{d_i + d_{i+1}} \right| /(m-1)$$

The modification of the original formula was suggested by Gibbon and Gut (2001) and it does not change the patterns found in the results. As it is more convenient, it will be used in the current study. However, if a comparison of results is required, the values obtained using the older formula on the left must be halved. (To avoid confusion, Gibbon and Gut proposed a different name for their adjusted metric. They wanted to call it the Rhythm Ratio. For the reasons stated in the Introduction, this proposal will be ignored.)

One of the metrics suggested by Ramus, Nespor and Mehler (1999) is a well-known and commonly used indicator of variation – the standard deviation from the mean. The authors showed that it was especially useful if calculated for durations of consonantal intervals, i.e., stretches of speech between two vowels filled with one or more consonants. They labelled it Δ*C*. The general formula can be found in most textbooks of statistics. For our purpose, it would be:

$$\Delta C = \sqrt{\frac{\sum_{i=1}^{m=1}\left(d_{Ci} - \bar{d}_C\right)^2}{m-1}}$$

where *i* is the summation index, *m* is the number of consonantal intervals in a breath-group, d_{Ci} is the duration of an i-th consonantal interval, which has to be subtracted from the mean duration of all consonantal intervals. The measure ΔV can be calculated analogically.

The resulting Δ*C* or ΔV is in milliseconds and it is quite sensitive to articulation rate. Therefore, it can be normalized into a coefficient of variation (as suggested by, e.g., Dellwo & Wagner, 2003) and commonly used under the name of Varco. It is a unit-less ratio, conventionally conceptualized as a percentage, but not necessarily written with the percentage symbol. The following formulae are for consonants and vowels respectively (the denominators are mean durations of consonantal or vocalic intervals):

$$VarcoC = 100 \times \frac{\Delta C}{\bar{d}_C} \qquad \text{or} \qquad VarcoV = 100 \times \frac{\Delta V}{\bar{d}_V}$$

Conceptually the simplest is the proportion of vocalic stretches in an utterance (or in our case in a breath-group), known as %V. It is calculated with the following formula:

$$\%V = 100 \times \frac{\sum_{i=1}^{m=1} d_{Vi}}{d_{BG}}$$

where *i* is the summation index, *m* is the number of vocalic intervals in the given breath-group, and d_V is the duration of a vocalic interval, while d_{BG} is the duration of the investigated breath-group. The result is expressed as a percentage.

2.4 Procedure

The metrics were computed for each breath-group (about 50 BGs per speaker) and weighted arithmetic means were calculated for random fifths of the breath-groups by a speaker. Hence, longer breath-groups contributed to the mean more than shorter one. (Weighting was based on the number of syllables in a breath-group, not the duration in seconds. Breath-groups of fewer than 5 syllables were ignored altogether.) The random fifths provide 5 instances of resulting values per news bulletin and are used to consider intra-speaker consistency.

Apart from raw measurements in the speech material as a whole, the exclusion of phrase-final portions was carried out to establish its influence on the results. Phrase final lengthening is a quasi-universal prosodic feature that poses a potential problem to the calculation of the metrics. Its domain might be unstable (anything from the last phone to the last stress-group), but, more importantly, its scale varies hugely. We thus repeated all measurements on the material will all phrase-final words excluded.

Since fluctuations in articulation rate are also reported phrase-initially (for initial acceleration see, e.g., Byrd & Saltzman, 2003 or Volín & Skarnitzl, 2007), on unusual or foreign words and during hesitation, one half of the speech material was cleansed to establish the influence of the aforementioned phenomena on the values of the metrics. The names of foreign politicians, cities and countries including their derivations (e.g., *barmský*, i.e., *Burmese*), words with hesitations in their pronunciation together with final and initial two-syllable stretches were excluded from the third round of measurements.

3. Results

3.1 Artificial Data

As explained above in Section 2, *durational variation metrics* were measured in artificial material simulating an articulation rate of 5 syll/s. Mutual ratios of vowel durations were manipulated to establish the changes in metrics values for given ratios. Since the given ratios would produce the same result if they were conceptualized for consonants, we will report these results for a general Segment (S), hence, %*S*, *nPVI-S*, Δ*S* and *VarcoS*. The metric %*S* was actually kept constant at the value of 50. Table 2 presents the results for *nPVI-S*, Δ*S* and *VarcoS*.

Table 2. Values of global temporal metrics for varying mutual ratios of units measured. The *VarcoS – short* relates to stretches of 1.6 s, while *VarcoS – long* was measured in stretches of 6.4 s. The column *ΔS* presents mean across four measurements on stretches of varying length.

S1 : S2 ratio	AR (syll/s)	nPVI-S	ΔS	VarcoS – short	VarcoS – long
1 : 1.25	5	11.1	11.5	11.9	11.3
1 : 1.5	5	20.0	20.7	21.4	20.3
1 : 2	5	33.3	34.5	35.6	33.8
1 : 3	5	50.0	51.7	53.5	50.8
1 : 4	5	60.0	62.1	64.1	61.0
1 : 5	5	66.7	69.0	71.3	67.8
1 : 6	5	71.4	73.9	76.3	72.5
1 : 7	5	75.0	77.6	80.2	76.7
1 : 8	5	77.8	80.5	83.2	79.0
1 : 9	5	80.0	82.8	85.5	81.3
1 : 10	5	81.8	84.7	87.4	83.1
1 : 19	5	90.0	93.1	96.2	91.4
1 : 39	5	95.0	98.3	101.6	96.5
1 : 2	3	33.3	59.3	38.4	34.4
1 : 2	8	33.3	21.1	34.8	33.7

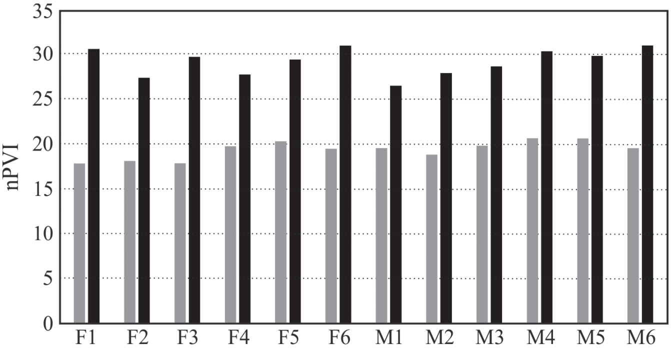
An important thing to note is that *nPVI* is not increasing linearly with the increase of the ratios. That is not surprising from the mathematical point of view, but when people conceptualize their results they should be aware of this. Another mathematically trivial fact is that *ΔS* and *VarcoS* are equal (apart from the units in which they are expressed), if the mean duration is 100. It is useful to notice that Table 2 contains the ratio of 1 : 2 three times for three different articulation rates. While *nPVI* is not affected by the changes in AR at all, *ΔS* changes dramatically (see the last two lines in the table) and *VarcoS* somehow normalizes, even if not perfectly.

The Pearson correlation between *nPVI-S* and *VarcoS* in our simulations was almost perfect: $r = 0.999$.

3.2 Natural Data

The *nPVI* values ranged from 17.8 to 20.7 for vowels and from 26.5 to 31.0 for consonants. Figure 2 demonstrates that the dispersion of individual speakers' values is relatively narrow.

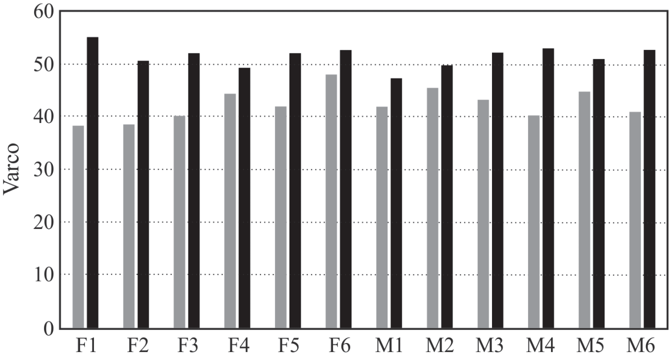
Figure 2. Values of nPVI for individual speakers. Black columns represent consonants, grey columns represent vowels. (For comparison with the older Grabe-Low procedure, our values would have to be doubled – see *Method*.)



It can be observed that consonantal pairwise variation is greater than the vocalic one. Consonant clustering is relatively common in Czech. The existence of phonological length in the vowel system does not seem to have any particularly profound influence on variation in vowel durations. In comparison with Dankovičová & Dellwo (2007) who also worked with Czech material, our *nPVI-V* values are lower (their mean was 23, ours is 19.47) and they are further lowered if the phrase-final lengthening is excluded (see below). Consonantal values cannot be compared since only raw (non-normalized) values are reported in Dankovičová & Dellwo (2007). The mutual correlation between consonantal and vocalic *nPVI* was established at $r = 0.11$ and was not statistically significant, which means that consonantal and vocalic measures vary independently. This fact is also hinted at by speaker F1 in Figure 2, who has the lowest *nPVI-V*, but one of the highest *nPVI-C*.

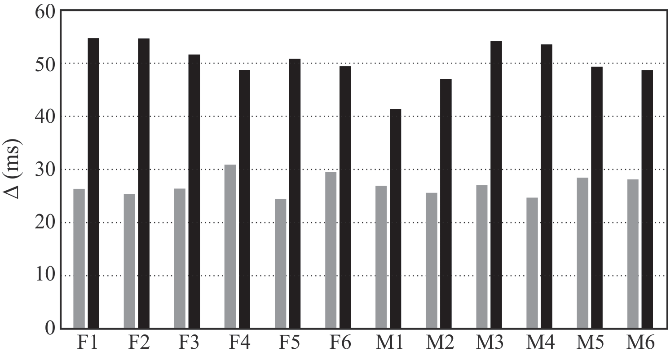
The values of *Varco* seem to be less compact than the previous metric: they ranged between 38.4 and 48.1 for vowels and between 47.4 to 55.2 for consonants (see Figure 3). Their insignificant mutual correlation also confirmed independence of vocalic and consonantal variation. However, the correlation between *nPVI-V* and *Varco-V* was established at $r = 0.55$ ($p < 0.001$) and for consonants (*nPVI-C* × *Varco-C*) even $r = 0.76$ ($p < 0.001$). Although this is far from almost perfect correlation found in the artificial data, there is still quite a high common trend between the two metrics. On the other hand, the difference in the behaviour of vowels and consonants speaks against attempts to replace one measure with the other. Dankovičová and Dellwo did not report *Varco-V* value for their sample, but their *Varco-C* was about 61, which exceeds even the highest value achieved by speaker F1, let alone our mean of 51.5 (Dankovičová & Dellwo, 2007).

Figure 3. Values of *Varco* (coefficient of variation) for individual speakers. Black columns represent consonants, grey columns represent vowels.



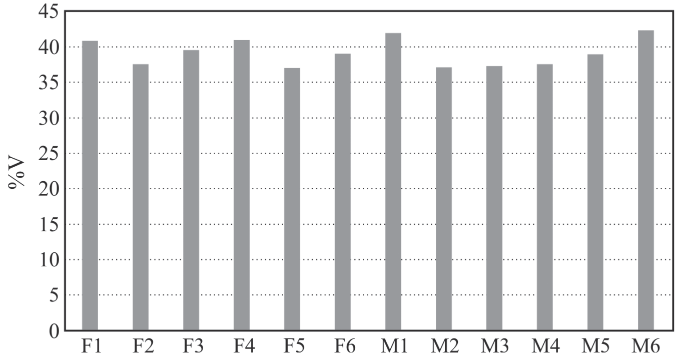
Non-normalized counterparts of *Varco* measures are standard deviations from mean durations labelled Δ after Ramus et al. (1999). They are displayed in Figure 4. The ΔV values range from 24.4 to 30.9 ms, whereas the ΔC values from 41.1 to 54.7 ms. The correlation coefficient for ΔV against *VarcoV* was established at $r = 0.79$ ($p < 0.001$) and for ΔC against *VarcoC* at $r = 0.75$ ($p < 0.001$).

Figure 4. Values of ΔV and ΔC (standard deviations from mean durations) for individual speakers. Black columns represent consonants, grey columns represent vowels.



The last reported metric is the percentage of vowel durations in the duration of a given stretch of speech. It is reported without its consonantal counterpart since it is measured in speech without pauses. Therefore, the consonantal measure would be perfectly correlated, i.e., $\%V + \%C = 100$. Figure 5 shows that $\%V$ is again quite similar across the speakers in the sample: the values range between 37.1 and 42.3%. Interestingly, the value reported by Dankovičová and Dellwo (2007) is again incompatible: their sample mean was over 46%.

Figure 5. Values of the metric $\%V$ (percentage of vowel durations in utterances) for individual speakers.



We did not expect this measure to correlate with any other metric, but the Pearson coefficients were calculated anyway. Indeed, there were no statistically significant trends between $\%V$ and either of *nPVI-V*, *nPVI-C*, *VarcoV* or *VarcoC*. Surprisingly, though, a negative correlation was found with ΔC ($r = -0.47$; $p < 0.001$) and a positive one with ΔV ($r = 0.46$; $p < 0.01$). This suggests that speakers with greater variation in consonant durations have lower proportion of vowels in their speech, while speakers with greater variation in vowel durations have the opposite. This effect disappears once the standard deviation is normalized by the mean.

3.2.1 Local changes in tempo

It is generally known that prosodic phrases are the domain of articulation rate change. Especially the phrase-final lengthening is one of the quasi-universals in the languages of the world. It has also been measured in Czech (Dankovičová, 2001; Volín & Skarnitzl, 2007).

Table 3. Differences in durational variation metrics in all speech material (All), after exclusion of the phrase-final word (W/O Final) and after cleansing in line with the conditions set in Method (Cleansed).

Metric	nPVI-V	nPVI-C	VarcoV	VarcoC	ΔV	ΔC	%V
All	19.47	29.29	42.22	51.63	26.80	50.60	39.07
W/O Final	18.81	28.92	36.42	50.65	21.26	46.59	37.75
Cleansed	18.07	29.14	37.00	51.05	21.70	47.38	38.73

The figures in Table 3 clearly indicate that the exclusion of the final word leads to a decrease in variation for all the metrics. This change is greater for vowels than for consonants. However, excluding initial and final two syllables in each phrase, foreign names and hesitations (the *cleansed* condition) does not strengthen this trend. Only *nPVI-V* decreases further, while the values of other metrics slightly rise again, even if not back to the values for the complete material. In other words, the influence of word-final lengthening is obvious, while fluctuations in tempo at the beginnings of phrases and in unusual words do not seem to have a clear effect in the speech style that was investigated here.

Figure 6. Scatterplots of *nPVI* values of randomly paired speakers, calculated for random fifths of their spoken texts. Male speakers are represented by empty circles, female speakers by filled squares.

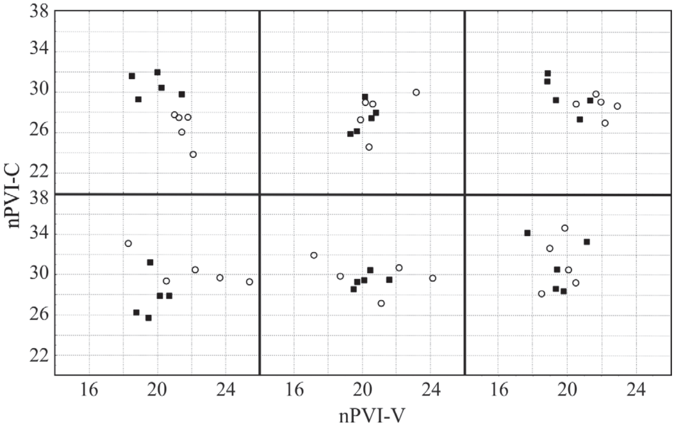
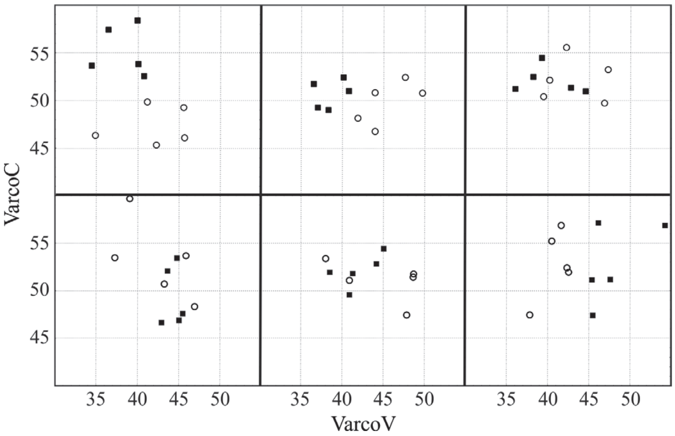


Figure 7. Scatterplots of *Varco* values of randomly paired speakers, calculated for random fifths of their spoken texts. Male speakers are represented by empty circles, female speakers by filled squares.



3.2.2 Intraspeaker variation

The results depicted in the Figures 2 to 5 above show that the interspeaker variation is not very high, but we felt it necessary to establish the magnitude of intraspeaker variations as well. In Figures 6 and 7 the speakers were randomly paired (by alphabetic cues from their surnames). In each of the scatterplots there are nPVI or Varco values of a male and a female speaker. As explained in the Method, the five values represent five random fifths of their breath-groups, i.e., about 100 words of the spoken text.

It is obvious that the distance between one speaker's values is seldom smaller than the distance between different speakers' values. In Figure 6 there is only one case (the top left scatterplot) in which a line could be drawn between the values of the two speakers. In Figure 7 this could be done for three pairs. Yet clearly, the dispersion of data within a speaker is comparable to the dispersion between speakers.

For two of the speakers (one male and one female) an additional news bulletin was obtained. It was recorded several weeks after the first recording. The format is identical with the first recording, but because it is a genuine instance of news reading broadcast by a national radio station, the text differs. The values of DVMs are presented in Table 4.

Table 4. Differences in durational variation metrics between two independent recordings (*a* and *b*) for two speakers (F4 and M4).

Speaker	nPVI-V	nPVI-C	VarcoV	VarcoC	ΔV	ΔC	%V
F4a	19.7	27.8	44.4	49.3	30.9	48.7	41.0
F4b	18.8	27.7	37.8	52.1	26.3	50.2	41.8
M4a	22.0	30.4	42.5	53.2	26.2	53.2	38.1
M4b	19.4	30.4	38.0	52.7	23.2	54.0	37.1

The most stable measure seems to be that of *nPVI-C*, while the vocalic measures fluctuate. Paradoxically, for *ΔV* the second reading of the female speaker and the first reading of the male speaker led to almost the same values, while their other readings differed notably. We can conclude that even quite an extensive spoken text (about 500 words) does not completely stabilize values of DVMs. This speaks against putting too much emphasis on "exact" numbers, and also against the use of small samples and subsequent generalizations for the whole language.

3.2.3 Czech versus English

The values retrieved for Czech speakers were compared with analogous data from English. News bulletins of the BBC World Service are of a very similar format (read monologues of about 500 words) and recordings of 4 women and 4 men speaking Southern British Standard were used. These figures are taken from the study Słówik & Volín (in print). The methodology of material processing was identical to the procedure used in the current study.

Table 5. Differences in durational variation metrics between Czech and English read monologues processed in a uniform manner.

Metric	nPVI-V	nPVI-C	VarcoV	VarcoC	ΔV	ΔC	%V
Czech	19.5	29.3	42.2	51.6	26.8	50.6	39.1
English	37.2	34.6	58.5	53.2	46.3	59.0	40.6

Table 5 shows that apart from *%V*, which is very similar for both language samples, there are clear differences in the rest of the measures. These are especially large in vowels:

nPVI-V is almost doubled in English. Despite the presence of phonological length in the Czech vowel system, the variation in durations of vowels is much smaller in comparison with English. This could be the consequence of the absence of unstressed reduced vowels in Czech, but also of the fact that text frequencies of long vowels in Czech are relatively low (only one in five vowels in texts is long). In addition, the Czech lexical stress does not lead to increased durations of vowels.

The difference in consonantal variation is smaller, yet not negligible. The explanation could perhaps be sought in phonotactics. Both languages allow for consonant clustering in syllable onsets and codas, but again, although coda clusters are possible in Czech, they are of low text frequencies.

4. Discussion

If rhythm is defined as a specific alternation of prominence patterns, then it should be viewed as multidimensional. Prominences arise from delicate interactions of F0 changes, durations, intensities and spectral properties. Pure durational measurements can hardly lead to comprehensive rhythm modelling. However, capturing durational variation in speech is still a necessary step towards complex models of speech rhythm. Rather than scandalising the durational variation metrics (previously also labelled as rhythm metrics), we should criticise their misinterpretation. Similarly, rather than denying the existence of rhythm in speech just because it is not neat enough, we should concentrate our effort on the description of prominence configurations typical of individual languages and speaking styles.

Alternatively, the research might perhaps re-focus on the flow of speech as such, which assumes some regularity and predictability of configurations, but is not as tightly linked to notions of monotony and simplicity. The motivation in most aspects would remain the same. Apart from arguments already stated above in the Introduction, we could evoke an old yet inspiring study by Miller and Hewgill (1964), who examined the correlations between dysfluent speech and credibility ratings, and found an inverse relationship: the more dysfluencies there are in speech, the lower the credibility ratings. After all, the etymology of the word rhythm shows clear link to the concept of flow.

The results presented here are coherent – they do not comprise randomly dispersed values. It follows that durational variation metrics or DVMs reflect certain properties of temporal organization that might play a useful role in speech research. Comparison of natural data with values achieved in simulations could perhaps inspire further considerations in this research field. The fact that representative English and Czech samples processed by identical methods mutually differ and can be related to the table of simulations also opens room for further thought.

On the other hand, it should also be noted that the results achieved in our study are in disagreement with another study (mentioned above) that mapped a Czech sample. The nature of the disagreement is difficult to clarify since the material of the previous study is insufficiently described – there are no specifications of its extent and circumstances of its collection. Too many published accounts base their findings on five sentences per language (sic!) or some unspecified material (e.g., "three subjects were asked to read a passage"). Current research community will quite certainly agree that such practice may bring mistrust to the research field and should, therefore, become a thing of the past.

Finally, although DVMs were defended in the present study, we should be ready to accept that these relatively crude measures of variation are not as useful as originally assumed and that they have to be replaced with better research tools. If that happens, the metrics should still be acknowledged as elements that paved the path to new discoveries.

ACKNOWLEDGEMENT

This research was supported by the Charles University project *Progres 4, Language in the Shiftings of Time, Space, and Culture*.

REFERENCES

Barry, W., Andreeva, B. & Koreman, J. (2009). Do rhythm measures reflect perceived rhythm? *Phonetica*, 66, 78–94.

Brugos, A. & Barnes, J. (2014). Effects of dynamic pitch and relative scaling on the perception of duration and prosodic grouping in American English. In: *Proceedings of 7th Speech Prosody*.

Buxton, H. (1983). Temporal predictability in the perception of English speech. In: Cutler, A. & Ladd, D. R. (Eds.), *Prosody: Models and Measurements*, 111–121. Berlin: Springer-Verlag.

Byrd, D. & Saltzman, E. (2003). The elastic phrase: modelling the dynamics of boundary-adjacent lengthening, *Journal of Phonetics*, 31, 149–180.

Cumming, R. E. (2011). The language-specific interdependence of tonal and durational cues in perceived rhythmicality. *Phonetica*, 68, 1–25.

Dankovičová, J. (2001). *The Linguistic Basis of Articulation Rate Variation in Czech.* Frankfurt am Main: Hector (Forum Phoneticum 71).

Dankovičová, J. & Dellwo, V. (2007). Czech speech rhythm and the rhythm class hypothesis. In: *Proceedings of 16th ICPhS*, 1241–1244.

Dellwo, V. & Wagner, P. (2003). Relations between language rhythm and speech rate. In: *Proceedings of 15th ICPhS*, 471–474. Barcelona: UAB & IPA.

Ghitza, O. & Greenberg, S. (2009). On the possible role of brain rhythms in speech perception: Intelligibility of time-compressed speech with periodic and aperiodic insertions of silence. *Phonetica*, 66, 113–126.

Gibbon, D. & Gut, U. (2001). Measuring speech rhythm. In: *Proceedings of Eurospeech 2001*, 91–94.

Gorow, R. (2000). *Hearing and Writing Music*. California: September Publishing Studio.

Grabe, E. & Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. In: Gussenhoven, C. & Warner, N. (Eds.), *Papers in Laboratory Phonology 7*, 515–546. Berlin: Mouton de Gruyter.

Grossberg, S. (2003). Resonant neural dynamics of speech perception. *Journal of Phonetics*, 31, 423–445.

Hove, M. J. & Risen, J. L. (2009). It's all in the timing: Interpersonal synchrony increases affiliation. *Social Cognition*, 27(6), 949–961.

Huggins, A. W. F. (1979). Some effects on intelligibility of inappropriate temporal relations within speech units. In: *Proceedings of 9th ICPhS,* 283–289.

Kinoshita, N. & Sheppard, Ch. (2011). Validating acoustic measures of speech rhythm for second language acquisition. In: *Proceedings of 17th ICPhS*, 1086–1089.

Kirschner, S. & Tomasello, M. (2010). Joint music making promotes prosocial behaviour in 4-year-old children. *Evolution and Human Behaviour*, 31, 354–364.

Knight R. A. (2011). Assessing the temporal reliability of rhythm metrics. *Journal of International Phonetic Association*, 41, 271–281.

Knight, S. & Cross, I. (2012). Rhythms of persuasion: The perception of periodicity in oratory. In: *Book of Abstracts – Perspectives on Rhythm and Timing*, p. 27. Glasgow: University of Glasgow.

Kohler, K. (2009). Rhythm in speech and language. *Phonetica*, 66, 29–45.

Loukina, A., Kochanski, G., Rosner, B., Keane, E. & Shih, Ch. (2011). Rhythm measures and dimensions of durational variation in speech. *Journal of Acoustical Society of America*, 129/5, 3258–3270.

Low, E. L., Grabe, E. & Nolan, F. (2000). Quantitative characterisations of speech rhythm: Syllable timing in Singapore English. *Language and Speech*, 43, 377–401.

Low, E. L. & Grabe, E. (1995). Prosodic patterns in Singapore English. In: *Proceedings of 13th International Congress of Phonetic Sciences*, 636–639.

Mariano, P. & Romano, A. (2011). Rhythm metrics for 21 languages. In: *Proceedings of 17th ICPhS*, 1318–1321.

Miller, G. R. & Hewgill, M. A. (1964). The effect of variations in non-fluency on audience ratings of source credibility. *Quarterly Journal of Speech*, 50/1, 36–44.

Nolan, F. & Jeon, H.-S. (2014). Speech rhythm: a metaphor? *Philosophical Transactions of the Royal Society B*, 1–11.

O'Rourke, E. (2008). Speech rhythm variation in dialects of Spanish: applying the pairwise variability index and variation coefficients to Peruvian Spanish. In: *Speech Prosody 2008*, 431–434.

Pollák P., Volín, J. & Skarnitzl, R. (2007). HMM-based phonetic segmentation in Praat environment. In: *Proceedings of 12th Intern. Conf. Speech & Computer – SPECOM 2007*, 537–541, Moscow: MSLU.

Quené, H. & Port, R. F. (2005). Effects of timing regularity and metrical expectancy on spoken-word perception. *Phonetica*, 62, 1–13.

Ramus, F., Nespor, M. & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, 73, 265–292.

Skarnitzl, R. & Volín, J. (2012). Referenční hodnoty vokalických formantů pro mladé dospělé mluvčí standardní češtiny [Reference values of vowel formants for young adult speakers of Standard Czech]. *Akustické listy*, 18/1, 7–11.

Slówik, O. & Volín, J. (in print). Acoustic correlates of temporal structure in North-Vietnamese English. In: J. Volín & R. Skarnitzl. (Eds.), *The Pronunciation of English by Speakers of Other Languages*. Newcastle: Cambridge Scholars Publishing.

Volín, J. & Churaňová, E. (2010). Probabilities of consonantal sequences in continuous Czech texts. *AUC – Philologica 1, Phonetica Pragensia XII*, 49–62.

Volín, J. & Skarnitzl, R. (2007). Temporal downtrends in Czech read speech. In: *Proceedings of 8th Interspeech*, 442–445.

Volín, J., Skarnitzl, R. & Pollák, P. (2005). Confronting HMM-based phone labelling with human evaluation of speech production. In: *Proceedings of Interspeech 2005*, 1541–1544.

Wagner, P. & Dellwo, V. (2004). Introducing YARD and re-introducing isochrony to rhythm research. In: *Proceedings of Speech Prosody 2004*.

White, L., Mattys, S., Series, L. & Gage, S. (2007). Rhythm metrics predict rhythmic discrimination. In: *Proceedings of 16th ICPhS*, 1009–1012.

RESUMÉ

Článek si klade za cíl ukázat, že koreláty rytmických jazykových typů mohou být využívány při výzkumu řeči, pokud netrváme na tom, že odrážejí „řečový rytmus" a pokud řádně specifikujeme jejich užití. Studie pracuje s materiálem reálným i simulovaným a přináší reprezentativní hodnoty pro český a anglický materiál v několika modifikacích.

Jan Volín
Institute of Phonetics
Faculty of Arts, Charles University
jan.volin@ff.cuni.cz

2017 ACTA UNIVERSITATIS CAROLINAE PAG. 95–107
PHILOLOGICA 3 / PHONETICA PRAGENSIA

THE RELATION BETWEEN SUBJECTIVE AND OBJECTIVE ASSESSMENT OF SPEAKING RATE IN CZECH RADIO NEWSREADERS

JITKA VEROŇKOVÁ AND PETRA POUKAROVÁ

ABSTRACT

This article examined objective and subjective speaking rate and their relation. Read speech of 22 Czech radio newsreaders (13 male and 9 female) constituted the material for the study. The objective rate was measured as speech rate and articulation rate, both expressed in syllables per second. Two domains, a single news report and an intonation phrase, were chosen as the units of observation. A perception test was used to establish the subjective assessment of speaking rate. The test was made up of the full news reports and the subjects were asked to rate their tempo on a scale slow – normal – fast. The median of speech rate and articulation rate for the individual news reports was 5.7 syll/s and 6.2 syll/s, respectively. In general, the listeners rated the stimuli tempo as normal and there was no significant difference between the subjective evaluation of female and male speakers. The possible factors influencing the relation between the subjective and objective rating are discussed; no simple and direct relationship between them was found.

Key words: speech rate, articulation rate, intonation phrase, perception, Czech, media speech

1. Introduction

1.1 Speaking rate in Czech media speech

Radio and TV news broadcasts fall among read texts that are very often examined because newsreaders and other media speakers are taken as promoters of standard speech. Regarding speaking rate, there is objective evidence that speech pronounced in the Czech radio and TV has accelerated in the last decades. Mean speech rate of radio newsreaders until 1994 was 5.1 syll/s (males) and 4.8 syll/s (females), in 1996 it was 5.3 syll/s (males and females) (all measured by Bartošek, 2000)[1] and 6.2 syll/s in 2002 (according to Palková, published in Palková et al., 2003). Mean speech rate of Czech TV

[1] The values are rounded to one decimal place.

https://doi.org/10.14712/24646830.2017.36

weather forecast is 5.6 syll/s (Balkó, 1999). Researchers pointed out the fast speaking rate in the mass media and difficulties in speech intelligibility that might be caused just by it (Bartošek, 1995, Palková, 2004, Havlík et al., 2013).[2]

We decided to inquire into this phenomenon and to inspect the relation between the objective and subjective speaking rate in Czech news broadcasts, because in previous experiments researchers focused mainly on objective measurements without systematic inspection of the subjective perspective of listeners.

In the experiment, both the speech rate and articulation rate[3] of Czech radio newsreaders were measured and a perception test was created to receive listeners' assessment. Then the measured (objective) speaking rate was compared with the listeners' ratings, i.e. the perceived (subjective) speaking rate, to establish the degree of correlation.

1.2 Variability in speaking rate

Speaking rate shows both the variability between and within speakers; there are many factors that affect it, both extralinguistic and intralingustic. However, their influence on the speaking rate may not be direct; the factors may complement one another, but they may also be in contradiction (conf., e.g., Kohler, 1986).

Verhoeven et al. (2004) examining the Dutch corpus of semi-spontaneous speech found that age, sex and the dialect region affect speaking rate. According to their experiment younger people spoke faster than the older, men spoke faster than women and speakers from the West region, considered to be the linguistic centre of the Netherlands (sic!), showed higher speech rate compared to speakers from dialect regions further from the centre. Quené analysing the same corpus claimed that these factors (age, gender and centre/periphery) may play a role, however he considered the length of the intonation phrase to be the most important factor. He quoted findings of Nooteboom and Lindblom – Rapp from the 70s and explained that longer phrases containing more syllables showed a tendency towards a faster speaking rate and a shorter syllable duration.[4] According to Quené's (2005) experiment, younger people tended to use longer intonation phrases and this was the cause of their faster speaking rate. However, Quené admitted that it was not possible to provide such a direct explanation for the other factors. There were other aspects, arising from the linguistic structure of texts, that caused changes of speaking rate, such as syllable structure (Pfitzinger, 2006), or phrase-final deceleration (for Czech Dankovičová, 2001; Volín & Skarnitzl, 2007), etc.

1.3 The relation between objective and subjective speaking rate

The variability of speech rate is also reflected in perception. Just as in speech production, there are many factors that affect our perception of speech rate. Kohler (1986), investigating German words and sentences, found overall duration, overall F0 level and

[2] In informal discussions listeners (especially the elderly) also complain that speaking rate in Czech media is very fast.

[3] More details about these measurement in 1.4.

[4] Ondráčková (1954a, b) (among others) refers to this phenomenon on the level of stress unit in Czech.

F0 movements to be important tempo cues. He examined not only the strength of the single factors, but pointed out that there was a relation between them and also between the factors and the production patterns. J. Koreman (2006) confirmed that articulation rate played an important role, yet not an exclusive one, for the perception of speech rate; other factors, e.g. pausing, disfluencies and other prosodic properties of speech, also determined the perceived speech rate. In his experiment the slow and fast speakers differed in their objective speech rates, however the grouping of the speakers according to the perceived speech rate did not match the categorisation based on the measured speech rate.

H. Pfitzinger (1998), who focused on local speech rate, claimed that linear combination of both the syllable and the phone rates (see 1.4 bellow) corresponded best to the perception of speaking rate.

1.4 Measurement of speaking rate – methodological notes

Speaking rate tells us, in general, the number of speech units produced by a speaker over a time unit. It is often expressed in syllables per second (syllable rate) or in phones per second (phone rate). It could also be expressed as the average syllable/phone duration. The interval unit could also be defined as vocalic and consonantal intervals (VC intervals) (conf. BonnTempo-Corpus (Dellwo et al., 2004)).

The chosen domain used to establish the speech rate did crucially influence the obtained results; it especially affected the degree of the speech rate variability (conf. Miller et al. (1984)). In the present study two domains were applied: a single item of a news bulletin and a unit we call an intonation phrase. The reason to choose an intonation phrase was based on the research of Dankovičová (1997, 2001). She examined the variability of articulation rate in Czech comparing three domains (breath group, syntactic phrase and intonation phrase) and she found that it was only the intonation phrase that had regular patterns of articulation rate. This unit was phonetically defined as a group of stress units combined together into a compact intonation unit. However, to determine this unit in Czech the sound properties of its boundaries were crucial (Palková, 1997, Daneš, 1957). Janoušková (2008) formalized and experimentally verified the hierarchy system of sound units in Czech and showed that a pause, a distinct melodic contour and the presence of final lengthening were important boundary markers of an intonation phrase. See the Appendix for an example of stimuli with intonation phrase boundaries marked.

Pauses influence both the production and the perception of speaking rate. They were also taken into account in the procedure of speech rate measurements. In calculating speech rate, pauses were counted into the duration of speech; however, in articulation rate measurements the duration of pauses was excluded. The principal question was to define the minimal pause duration to determine an articulatory pause. The usual value ranged between 0.2 and 0.3 s, following the works of Goldman-Eisler (1961). Hieke et al. (1983) recommended a value closer to 0.1 s; the duration of 0.13 s was used by Dankovičová (2001).

2. Experiment

2.1 Material

For the purpose of the experiment news broadcasts of the public Czech Radio (CRo) were used. The recordings were obtained from the web archive of CRo[5] using the Cool-Edit96 software, sampling frequency 32 kHz, 16-bit amplitude resolution[6]; all the recordings were taken in the period between February–April 2008. The news bulletins were divided into single news reports; only the main speakers (i.e., the news readers) were taken into account, the speech of correspondents, analysts etc. was excluded as well as the news with backchannel sounds and disfluencies such as slips of the tongue. To eliminate yet another potential variable, the topic of the news was limited to "politics"; sports news and weather forecasts were also omitted. According to these criteria 22 speakers (13 male and 9 female) and their 22 news reports (1 recording per speaker) formed the corpus for the further procedure.[7]

2.2 Perception test

In order to keep the task as natural as possible, the news were presented to the listeners in continuous form, as retrieved from the source. The testing was carried out one year after retrieval of the material. Overall, the median length of the stimuli was 26.6 s, item length ranging between 24–32 seconds. Speech covered over 90.0% of each tested sample; except for one stimulus where the volume of pauses exceeded 10.0%. The stimuli contained 28 intonation phrases and 160 syllables on average and the number of intonation phrases/syllables was comparable in different stimuli. The intonation phrases that consisted of just one stress unit were the most numerous (46.3%).[8]

Each stimulus corresponded to one news report and it was played just once. Because of the total duration, the test was divided into two parts; each part lasted approx. 7 min., including 3 stimuli for the training session.

The listeners were Czech natives, students of the Czech language at the Faculty of Arts at Charles University in Prague, their mean age was 21 years. Seventeen listeners participated in both parts I and II (with an interval of 1 week in between the two parts); in addition, other 8 listeners participated in part II. All the listeners were females; a sufficient number of males was not available. The perception test was carried out in a sound treated lecture room. For each stimulus, the listeners judged the speaking rate of the speaker.

The listeners were presented with 3 main categories of speaking: slow (-1) – normal (0) – fast (1); they had the possibility to refine the evaluation using arrows to indicate a finer assessment within the basic category, in essence, it was a 9-point scale. For an illustration of the sheet form see Tab. 1 (part A). (Part B shows how the judgements were expressed numerically.)

[5] http://hledani.rozhlas.cz/iradio/, (recorder February–April 2008)

[6] Raw data: total duration 210 min., 24 newsreaders, total duration of news bulletins 4 min. on average.

[7] In addition 3 more news reports were chosen for the training session of the perception test.

[8] This finding is in accordance with Dankovičová (2001) and Hrachová (2016).

Table 1. Illustration of the answer sheet for the perception test (part A), with the corresponding numerical expression of the judgements (part B).

A					Speech Rate				
		Slow			Normal			Fast	
1.		-1			0			← [✓ 1]	
2.		-1			[✓ 0]			1	
...									
B	-1.33	-1.00	-0.66	-0.33	0.00	0.33	0.66	1.00	1.33

2.3 Segmentation of the speech material

First, each news report was transcribed orthographically. To prepare the canonical transcriptions, the software Convertor (Laun, 2001) and the template TRIPAC (Janoušková, 2003) were used and consequently manually corrected. The intonation phrases were labelled using the software Praat (Boersma & Weenink, 2002–2010). A pause was classified as an articulation pause if its duration was at least 0.2 s (see 1.4 above). Both the labelling and the listening method yielded the same results regarding the position of pauses.

The segmentation was done according to the rules based on Machač & Skarnitzl (2009). It was necessary to solve special cases where a consonant at the beginning of a word was preceded by a glottal gesture followed by a schwa. Such a preglottalized sequence was regarded as a part of the given word, and was counted as an extra syllable. (Conf. the findings about preglottalization in news reading in Skarnitzl & Machač, 2009.)

2.4 Laboratory measurements

Speaking rate was expressed in syllables per second, both for speech rate (SR) and articulation rate (AR). Except for preglottalization mentioned in 2.3, the canonical number of syllables in words was used to calculate speaking rate; in Czech, syllable compression was not frequent in general and it appeared only once in our material (unlike the elision of consonants).

The entire news report corresponding to one stimulus of the perception test served as the unit of observation; speaking rate was measured in the whole recording: the duration of each item was divided by the number of syllables to receive SR. The item duration reduced by the total duration of internal pauses and divided by the number of syllables served for the calculation of AR.

Because of the possible variation of speaking rate within the stimuli and its potential influence on the subjective assessment, articulation rate was also calculated in yet another way.

Another value of a speaker's articulation rate (ARIP) was obtained by determining the mean articulation rate of all intonation phrases within each recording (where the intonation phrase AR was calculated as the duration of a single intonation phrase divided by the number of syllables it contained).

The values of SR, AR and ARIP for each stimulus (N = 22) were obtained.

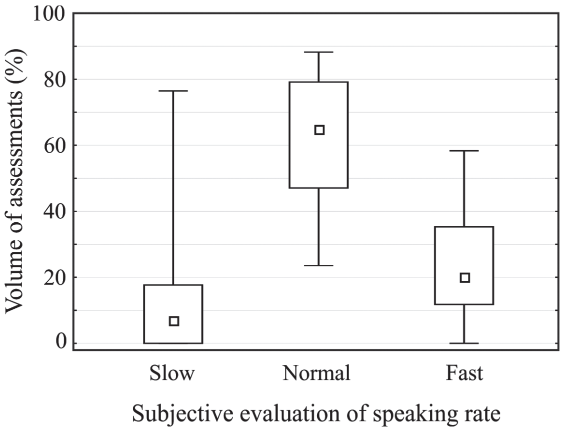
3. Results

3.1 Subjective speech rate

The speaking rate of every stimulus, i.e. each news report, was judged (see 2.3 above). Altogether there were 460 assessments available. The distribution of the global evaluation of speaking rate within the perception test was: slow (14.6%), normal (63.3%) and fast (22.2%). It is obvious that the listeners did not use all the categories equally in their assessments. The category normal speaking rate was chosen in two thirds of the assessments, while in almost one quarter of the cases the listeners assessed the speaking rate as fast.

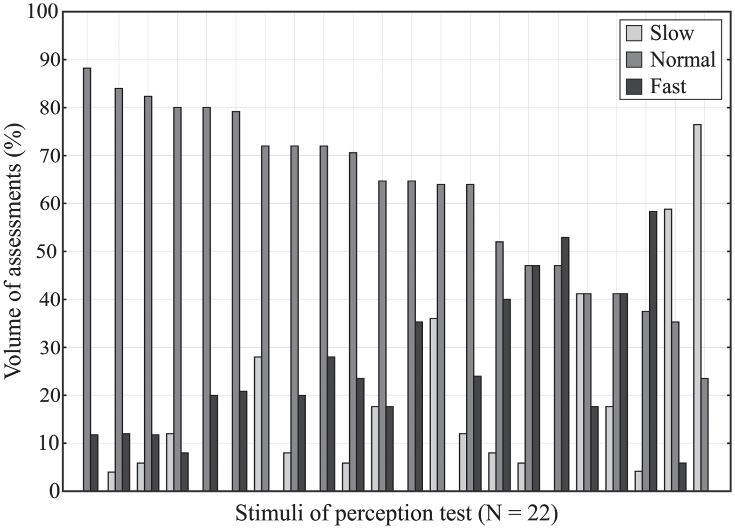
Subsequently the assessments of individual stimuli were examined. See Fig. 1 for the categories slow – normal – fast as a whole and Fig. 2 for individual stimuli.[9] The category normal speaking rate was not only chosen by the listeners most often but it also showed the highest agreement in the rating of single stimuli in comparison with the other categories. The range of agreement of rating the individual stimuli within the category normal was 23.5–88.2%, while the agreement on speaking rate in category slow was 0.0–76.5% and 0.0–58.3% for the category fast. It can be seen that at least 20% of the listeners agreed that the speaking rate of any given stimuli was normal, whereas the categories slow or fast remained unused in the evaluation of some stimuli.

Figure 1. Subjective evaluation of speaking rate. The volume of slow – normal – fast ratings (in %). Indicated are the median, quartile range, and range.



[9] All the tests, correlations and graphs were computed and visualized using software Statistica 12.0 (StatSoft).

Figure 2. Subjective evaluation of speaking rate of individual stimuli. The volume of slow – normal – fast ratings (in %).



For the following procedure, a recalculation of the subjective judgements was done (see Tab. 1 Part B) and the average evaluation of each stimulus was calculated. The tendencies shown above were confirmed: the speaking rate of stimuli was evaluated as normal on average.

According to the Mann-Whitney test there was no significant difference between the subjective evaluation of female and male speakers at alpha = 0.05 (the U-value is 30, the critical value of U at p < 0.05 is 28).

3.2 Objective speech and articulation rate

Table 2. Objective measurement: speech rate (SR), articulation rate (AR), articulation rate based on articulation rate of intonation phrases (ARIP). In syll/s.

syll/s	SR			AR			ARIP		
Female/Male	F	M	Total	F	M	Total	F	M	Total
Number of stimuli	9	13	22	9	13	22	9	13	22
Median	5.8	5.7	5.7	6.2	6.2	6.2	6.2	6.0	6.1
1st quartile	5.7	5.6	5.6	6.1	6.0	6.0	6.0	5.9	5.9
3rd quartile	6.0	5.9	6.0	6.4	6.5	6.5	6.3	6.3	6.3
Minimum	5.3	5.3	5.3	5.8	5.8	5.8	5.6	5.6	5.6
Maximum	6.2	6.2	6.2	6.6	6.9	6.9	6.5	6.8	6.8
Mean	5.8	5.8	5.8	6.2	6.3	6.3	6.2	6.1	6.1
Stand. dev.	0.3	0.3	0.3	0.3	0.4	0.3	0.3	0.4	0.3

For each stimulus, we measured speech rate (SR) and articulation rate (AR and ARIP)[10].

Median values of all the 22 stimuli (in syll/s) were: SR 5.7, AR 6.2, ARIP 6.1. The values for SR were lower than for the articulation rates, and there were visible overlaps between AR and ARIP. See Table 2 and the box plot in Fig. 3. The differences, tested by means of a t-test for correlated measures, were significant at alpha = 0.05 not only for SR and articulation rates, but also for AR and ARIP; SR – AR: t (21) = –17.4; SR – ARIP: t (21) = –10.9; AR – ARIP: t (21) = 4.5. According to the Pearson correlation coefficients all the correlations (SR – AR, SR – ARIP, AR – ARIP) were very high (r = 0.9 in all three cases). The scatter plots (Fig. 4a–c) show these relations very clearly.

Figure 3. Speech rate (SR), articulation rate AR and ARIP of stimuli (N = 22). Indicated are the median, quartile range, and range.

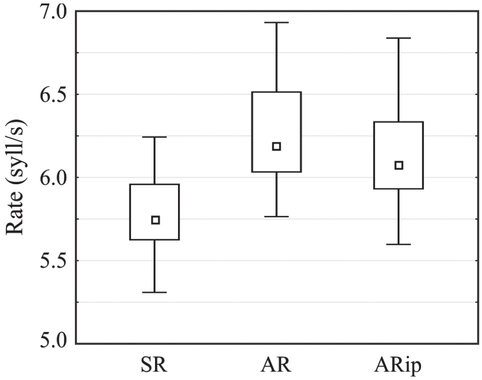
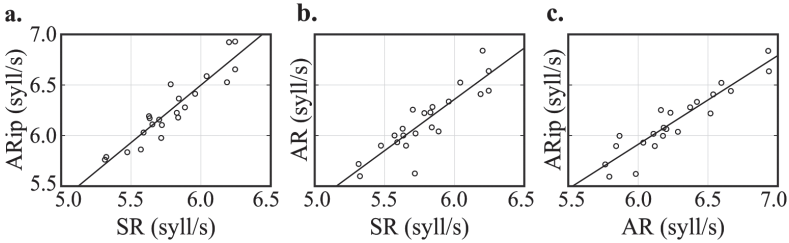


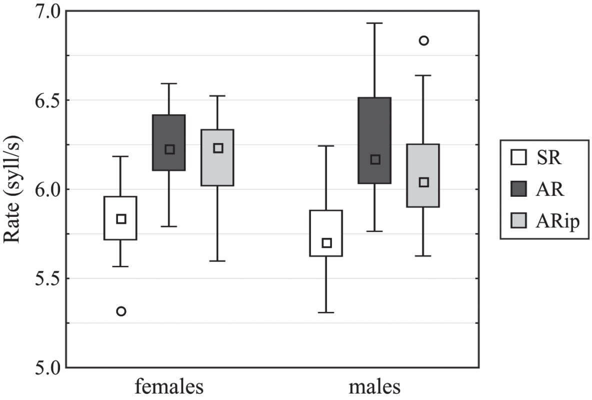
Figure 4. a. Articulation rate AR as a function of speech rate SR (N = 22), r = 0.93. **b.** Articulation rate ARIP as a function of articulation rate AR (N = 22), r = 0.93. **c.** Articulation rate ARIP as a function of speech rate SR (N = 22), r = 0.88.



The values of SR, AR and ARIP in female speakers were more compact than in males, but there were visible overlaps and according to the Mann-Whitney test, the differences between males and females in SR, AR and ARIP were not significant at alpha = 0.05. See Table 2 and box plot in Fig. 5.

[10] AR: articulation rate counted within the whole stimulus. ARIP: articulation rate counted as mean articulation rate of the intonation phrases. See 2.5 above.

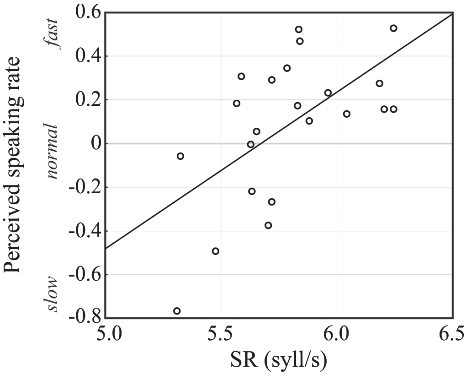
Figure 5. Speech rate (SR), articulation rate AR and ARIP of stimuli. Indicated are the median, quartile range, and range, as well as extreme values.



3.3 The relation between subjective and objective speech and articulation rates

Firstly, the relation between the subjective speaking rate and the measured values for SR, AR and ARIP was examined. Spearman rank correlation coefficients implied low/moderate correlations (rs = 0.5 in all three cases). The scatter plot in Fig. 6 shows the relation between SR and the subjective speaking rate.

Figure 6. Subjective speech rate as a function of speech rate SR (syll/s) (N = 22), r = 0.54



The influence of pauses, as another potential variable, was also examined. The Spearman rank correlation coefficient implied very low or low correlation between the amount of pauses and the examined parameters (SR: rs = 0.08, AR rs = 0.36, ARIP: rs = 0.30, subjective speaking rate: rs = 0.05).

It seems that there was no single correlate that would certainly describe the relation between the objective and the subjective evaluation of the tested stimuli in our corpus. This hypothesis could be illustrated by other specific examples. Speaker A was one of the objectively fastest speakers in our corpus (SR: 6.2 syll/s, AR 6.9 syll/s). Subjectively his speech rate was evaluated as fast by 52.9% of listeners and as normal by 47.1% of listeners. Speaker B received similar subjective evaluation (fast in 58.3% and normal in 37.5%) but had lower objective values (SR: 5.8 syll/s, AR 6.2 syll/s). On the other hand, subjective speech rate of speaker C was evaluated as normal in 88.2%, although the objective values were higher (SR: 6.0 syll/s, AR 6.6 syll/s).

4. Discussion

The speaking rate of read radio bulletins was examined from both the subjective and objective point of view.

Regarding the objective rate, both the speech rate (SR) and the articulation rate (AR and ARIP) were calculated. All of these parameters were highly correlated with each other and at the same time the differences between them were significant. On the other hand, there was no significant difference found between male and female speakers; this finding was in accordance with the results of Veroňková and Janoušková (partially published in Veroňková, 2012), who examined the speaking rate of TV newsreaders.

The values measured in our experiment (median values in syll/s SR 5.7, AR 6.2, ARIP 6.1) corresponded to SR of Czech newsreaders obtained by other researchers in their studies from the last decade: 6.2 syll/s measured by Palková (in Palková et al., 2003) or 5.9 syll/s (males) and 6.1 syll/s (females) (Veroňková & Janoušková)[11]. According to Palková (2004), the acceptable SR of news in Czech ranged between 5.5 and 5.8 syll/s. As far as our recordings are concerned, the SR of the tested news reports falled mostly to the higher end of this range, some even exceeded the limit (compare with a similar finding in Shevchenko & Uglova (2006) for TV news in the USA). Speaking rate of news presentation in professional speakers was consistently faster than SR in other genres or in non-professional speakers: the average SR of read speech pronounced by Czech university students was 4.7 syll/s (Balkó, 1999), 4.5 syll/s (Veroňková-Janíková, 2004); SR of guests performing in the radio broadcasts was 4.3 syll/s (Bartošek, 2000); SR of direct sport reports was 5.6 syll/s (ibid).

Speaking rate of the tested speech samples, i.e. news reports, was subjectively judged as normal on average. The same finding was examined by the above-mentioned study (Veroňková, 2012). Two factors could serve as an explanation for the low/moderate correlations between subjective judgement and objective measures of speech rate.[12] Firstly, the age of listeners may play a role. In both cases, the subjects were university students, i.e. younger adults. It is possible that the younger generation was more tolerant to higher speaking rates.[13] This opinion can be supported by the findings of Verhoeven et al. (2004) and Quené (2005) (mentioned in 1.1 above) regarding the faster SR of younger

[11] The data are based on the material of TV news from 2003 published partially in Veroňková (2012).

[12] One of the anonymous reviewers of this paper also pointed out a third factor: "The fact that most people mostly chose "normal" as the response ... [is] affected by the response options they had. Choosing the medium option is a frequent response strategy."

[13] According to the findings of Veroňková & Janoušková (2010) based on the material of TV news, younger people praise the (relatively) slower SR of newsreaders, however they do not complain about the objectively fast one.

people, together with the results of Schwab (2011) who confirmed the influence of the listeners' own objective SR to their subjective judgement of speaking rate. Secondly, the listeners probably took the type of communication and text into account – the SR within the genre, i.e. the news reporting, was subjectively being rated as normal. An experiment examining the subjective assessment of speech samples from different genres is being prepared.

In our experiment, which was focused on relatively long stimuli, no simple and direct relation between the subjective and objective rating was found. As it was discussed, there are many factors that could influence both the perception and the production of speech rate. It is clear that the longer the speech sample the harder it is to establish and mark off the one and most important factor as all of them are tightly bound together. There are several factors that deserve special attention in the following analyses of the corpora: F0 contour, pauses and the way of segmentation in general, the precision of articulation and phone rate. In the future studies, it would be useful to obtain the subjective speaking rate assessment of single intonation phrases from our corpora and to compare it both with the subjective rating of the given speaker and with the objective measurements.

REFERENCES

Balkó, I. (1999). K fonetickému výzkumu tempa řeči a tempa artikulace v čteném textu a spontánním projevu. In: Jinakost, cizost v jazyce a literatuře, (pp. 38–44). Ústí nad Labem: UJEP.

Bartošek, J. (1995). Jazyková kultura mluveného zpravodajství. In: J. Jančáková, M. Komárek & O. Uličný (Eds.), Spisovná čeština a jazyková kultura. Sborník z olomoucké konference 23.– 27. 8. 1993, Vol. 1 (pp. 166–171). Praha: FF UK.

Bartošek, J. (2000): Mluvní tempo v rozhlase a v televizi. In: Psychotrofon 1 (pp. 78–82). Olomouc: Univerzita Palackého.

Boersma, P. & Weenink, D. (2002–2010). Praat: doing phonetics by computer. (Version 5.0.10, retrieved February 27th, 2008 and 5.1.23, retrieved January 1st, 2010), http://www.praat.org.

Daneš, F. (1957). Intonace a věta ve spisovné češtině. Praha: Academia.

Dankovičová, J. (1997). The domain of articulation rate variation in Czech. *Journal of Phonetics*, 25, 287–312.

Dankovičová, J. (2001). The Linguistic Basis of Articulation Rate Variation in Czech. In: H. W. Wodarz (Ed.), *Forum Phoneticum 71*. Frankfurt am Main: Hector.

Dellwo, V., Steiner, I., Aschenberner, B., Dankovičová, J. & Wagner, P. (2004). BonnTempo-Corpus and BonnTempo-Tools: A database for the study of speech rhythm and rate. In: Proceedings of the 8th ICSLP (pp. 777–780).

Goldman-Eisler, F. (1961). The Significance of Changes in the Rate of Articulation. *Language and Speech*, 4, 171.

Havlík, M., Jílková, L., Kaderka, P. & Mrázková, K. (2013). Analýza a hodnocení jazykové úrovně vybraných pořadů České televize za 1. pololetí 2013. Retrieved February 2016 http://www.ceskatelevize.cz

Hieke, A. E., Kowal, S. & O'Connell, D. C. (1983). The trouble with „articulatory" pauses. *Language and Speech*, 26, 493–511.

Hrachová, J. (2016). Temporální charakteristika promluvových úseků v řeči profesionálních mluvčích. Unpublished diploma thesis. Prague: Institute of Phonetics, Faculty of Arts, Charles University.

Janoušková, J. (2003). TRIPAC. Originální aplikace šablony Microsoft Word pro potřeby fonetické transkripce češtiny. Interní nástroj Fonetického ústavu Filozofické fakulty univerzity Karlovy. Praha: Fonetický ústav FF UK v Praze.

Janoušková, J. (2008). Shoda percepčního hodnocení hloubky prozodických předělů v závislosti na struktuře čteného textu. In: J. Volín & J. Janoušková (Eds.), *AUC Philologica 1, Phonetica Pragensia XI* (pp. 87–104). Praha: Karolinum.

Kohler, K. J. (1986). Parameters of speech rate perception in German words and sentences: duration, F0 movement, and F0 level. *Language and Speech*, 29, 115–140.

Koreman, J. (2006). The Role of Articulation Rate in distinguishing fast and slow speakers. Proceedings of the 3rd International Conference on Speech Prosody. Retrieved December 12th, 2009, http://www.hf.ntnu.no/isk/koreman/Publications/2006/SpeechProsody2006.pdf

Laun, M. (2001). Convertor. Version 1.0.4. Software pro fonetickou transkripci češtiny.

Machač, P. & Skarnitzl, R. (2009). *Fonetická segmentace hlásek*. Praha: Nakladatelství Epocha.

Miller, J. L., Grosjean, F. & Lomanto, C. (1984). Articulation Rate and Its Variability in Spontaneous Speech: A Reanalysis and Some Implications. *Phonetica*, 41, 215–225.

Ondráčková, J. (1954a). O mluvním rytmu v češtině. I. Pojetí mluvního taktu. *Slovo a slovesnost*, 15, 24–29.

Ondráčková, J. (1954b). O mluvním rytmu v češtině. II. Členění mluvy. *Slovo a slovesnost*, 15, 145–157.

Palková, Z. (1997). *Fonetika a fonologie češtiny*. 2nd edition. Praha: Karolinum.

Palková, Z. (2004). Srozumitelnost řeči v rozhlasovém a televizním zpravodajství. In: Přednášky z XLVII. běhu Letní školy slovanských studií (pp. 97–109). FF UK: Praha.

Palková, Z., Veroňková-Janíková J. & Hedbávná, B. (2003). Zvuková podoba rozhlasové češtiny. In: Proměna rozhlasového výrazu a tvaru. Sborník příspěvků z jarního semináře (pp. 20–39). Praha: Sdružení pro rozhlasovou tvorbu.

Pfitzinger, H. R. (1998). Local Speech Rate as a Combination of Syllable and Phone Rate. In: Proceedings of ICSLP, Vol. 3 (pp.1087–1090). Sydney.

Pfitzinger, H. R. & Tamashima, M. (2006). Comparing Perceptual Local Speech Rate of German and Japanese Speech. In: Proceedings of the 3rd International Conference on Speech Prosody (ICSP 2006), Vol. 1 (pp.105–108). Dresden.

Quené, H. (2005). Modelling of Between-Speaker and Within-Speaker Variation in Spontaneous Speech Tempo. In: Proceedings of Interspeech (pp. 2457–2460). Lisbon.

Schwab, S. (2011). Relationship between speech rate perceived and produced by the listener. *Phonetica*, 68, 243–255.

Shevchenko, T. & Uglova, N. 2006. Timing in news and weather forecasts: implications for perception. Proceedings of the 3rd International Conference on Speech Prosody (ICSP 2006), Vol. 1 (pp. 5–8). Dresden.

Skarnitzl, R. & Machač, P. (2009). Domain-initial coordination of phonation and articulation in Czech radio speech. *AUC Philologica 1, Phonetica Pragensia XII*, 21–35.

Verhoeven, J., De Pauw, G. & Kloots, H. (2004). Speech Rate in a Pluricentric Language: A Comparison Between Dutch in Belgium and the Netherlands. *Language and Speech*, 47, 297–308.

Veroňková, J. (2012). Tempo řeči z různých stran. In: K. Hlínová & R. Vacula (Eds.), Sborník Asociace učitelů češtiny jako cizího jazyka. Praha: Akropolis, 203–223.

Veroňková-Janíková, J. (2004). Dependence of individual speaking rate on speech task. In: Z. Palková & J. Veroňková (Eds.), AUC – Philologica 1, Phonetica Pragensia X (pp. 107–123). Charles University in Prague: The Karolinum Press.

Veroňková, J. & Janoušková, J. (2010). Jak hodnotí posluchači zvukovou stránku projevů u moderátorů zpravodajství ČT1? In: S. Čmejrková, J. Hoffmannová & E. Havlová (Eds.), Užívání a prožívání jazyka (pp. 115–119). Praha: Karolinum.

Volín, J. & Skarnitzl, R. (2007). Temporal downtrends in Czech read speech. In: Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech 2007) (pp. 442–445). Antwerpen: ISCA.

http://hledani.rozhlas.cz/iradio/ (retrieved February–April 2008)

APPENDIX

An example of a news report – stimulus of the perception test (orthographic form) showing intonation phrase boundaries (‖). (P) refers to the noticeable pause.

Posádka raketoplánu ‖ Endeavour ‖ má za sebou ‖ úspěšný ‖ pátý výstup ‖ do volného kosmického prostoru. ‖ (P) Během šesti hodin ‖ připevnila ‖ na mezinárodní vesmírnou stanici několikametrový ‖ nástavec ‖ s kamerou. ‖ (P) Ta bude vesmírným lodím sloužit ‖ ke kontrole ‖ tepelného štítu, ‖

aby se už neopakovala tragédie z roku ‖ dva tisíce ‖ tři, ‖ (P) kdy ‖ malá trhlina ‖ způsobila ‖ rozpad ‖ raketoplánu ‖ Columbia ‖ (P) a všech ‖ sedm členů posádky ‖ zahynulo. ‖ (P) Astronauti z Endeavouru teď ‖ odpočívají ‖ před úterním ‖ odletem zpět k Zemi. ‖ (P)

Translation

The crew of the Endeavour space shuttle has completed the fifth extra-vehicular activity. During the six-hour EVA, they attached a several-meter long extender with a camera. It will serve for space shuttle heat shield tests so that a disaster, similar to the one of space shuttle Columbia from 2003 when a small crack caused its break-up and all seven crew members died, would not happen again.

The astronauts in the Endeavour space shuttle are now resting before they will take off back to the Earth on Tuesday.

RESUMÉ

Vztah mezi subjektivním a objektivním mluvním tempem u moderátorů českého rozhlasového zpravodajství

Experiment zkoumá objektivní a subjektivní mluvní tempo a jejich vzájemný vztah. Materiál tvoří nahrávky zpráv Českého rozhlasu (22 mluvčích – 9 žen a 13 mužů), od každého mluvčího 1 zpráva o délce 24–32 s. Subjektivní hodnocení bylo získáno na základě percepčního testu, položky tvořily jednotlivé zprávy. Posluchači hodnotili vzorky na škále tempo pomalé – střední – rychlé). Pro každý vzorek bylo změřeno mluvní tempo celkové (MTC) a artikulační tempo (ve slabikách za sekundu). Artikulační tempo bylo počítáno dvojím způsobem; shodně s mluvním tempem celkovým v rámci položky (AT) a dále v rámci jednotlivých promluvových úseků (ATú). Naměřené hodnoty mediánu činí pro MTC 5,7 sl/s, AT 6,2 sl/s a ATú 6,1 sl/s. Mezi MTC, AT a ATú byla prokázána vysoká korelace; zároveň bylo ověřeno, že rozdíly mezi nimi jsou statisticky významné. Naopak mezi muži a ženami není ve zkoumaném vzorku rozdíl v tempu řeči významný. Posluchači hodnotí položky z hlediska tempa obecně jako střední. Na produkci i percepci tempa působí vzájemně mnoho faktorů. To se potvrdilo i v našem experimentu; v rámci proměnných (MTC, AT, ATú, objem pauz) nebyl nalezen samostatný korelát platný pro všechny vzorky korpusu, který by jednoznačně charakterizoval vzájemný vztah mezi objektivním měřením a percepčním hodnocením. V další fázi výzkumu bude vhodné zaměřit se na roli melodického průběhu, výslovnosti a hláskového tempa, hlubšímu zkoumání podrobit pauzy a členění obecně. Užitečné by bylo percepční hodnocení vzorků na úrovni promluvových úseků a jeho srovnání s hodnocením tempa příslušného mluvčího.

Jitka Veroňková
Institute of Phonetics
Faculty of Arts, Charles University
jitka.veronkova@ff.cuni.cz

Petra Poukarová
Institute of the Czech National Corpus
Faculty of Arts, Charles University

2017 ACTA UNIVERSITATIS CAROLINAE PAG. 109–120
PHILOLOGICA 3 / PHONETICA PRAGENSIA

THE EFFECT OF VOICE QUALITY ON HIRING DECISIONS

LEA TYLEČKOVÁ, ZUZANA PROKOPOVÁ,
RADEK SKARNITZL

ABSTRACT

This paper examines the effect of voice quality on hiring decisions. Considering voice quality an important tool in an individual's self-presentation in the job market, it may very well enhance his/her job prospects, while some voice qualities may affect employers' judgments in a negative way. Five men and five women were recorded reading four different utterances representing answers to job interviewers' questions in four different phonation guises: modal, breathy, creaky and pressed. 38 professional employment interviewers recorded the speakers' hireability and personality ratings (likeability, self-confidence and trustworthiness) on 7-point semantic differential scales based on the speakers' voice. The results revealed a significant effect of the phonation guises on the speakers' ratings with the modal voice being superior to the cluster of non-modal voices. Interestingly, the non-modal guises were evaluated in a very similar way, except for the self-confidence category with the breathy voice getting the lowest scores on the one hand and the pressed voice correlating with high self-confidence ratings on the other.

Key words: voice quality, phonation types, speaker's perception, hiring decisions, matched guise technique

1. Introduction

Efficient sharing of information is one of the most characteristic aspects of the current period; in the digital era, more and more emphasis is being put on an individual's ability to communicate effectively and to convey messages clearly and accurately both verbally and non-verbally. The importance of an individual's voice in everyday interpersonal communication can thus hardly be overlooked (Laver, 1980: 1; DeVito, 2016: 48). Considering the contemporary job market context, required educational qualifications and professional experience certainly do not represent the only decisive factors in the recruitment and selection process (DeVito, 2016: 24). It is the overall self-presentation of applicants at job interviews that seems to play a very important role when it comes to hiring decisions. In this respect, voice quality is considered an essential

https://doi.org/10.14712/24646830.2017.37

element enabling a job candidate to present himself/herself in the appropriate way to secure employment (e.g., Skills You Need, 2016).

Although there exist quite a lot of studies on human voice, investigation into voice quality and its social role has not attracted much attention until recently. This question has become of great interest not only to the academic community, but also to a wide range of professionals as well as to the media (Greer & Winters, 2015). However, to our best knowledge, not many researchers have examined the importance of voice quality within social interaction in the Czech context.

According to numerous scientific findings, voice quality does play a vital role in interpersonal communication, as it is a significant indicator of the speaker's physical, psychological and social characteristics (Laver, 1980: 1; Kreiman, Vanlancker-Sidtis & Gerratt, 2003; Moisik, 2012). The subject of the present paper is the empirical mapping of voice quality as one of the essential factors in the hiring decision process.

1.1. Voice quality and types of phonation

Defining voice quality in a clear-cut, satisfactory and generally acceptable way is a rather challenging task (Kreiman et al., 2003). This term tends to be used in various contexts, e.g. a professional singer approaches voice quality in a different way than a phonetician (Childers & Lee, 1991). Nonetheless, even speech scientists do not seem to refer to voice quality unequivocally. The total auditory impression of the characteristic colouring of an individual speaker's voice may be seen, in a broad sense, as the result of both laryngeal and supralaryngeal features, i.e., differences in phonatory settings and vocal tract resonance characteristics, respectively. In the narrower sense, voice quality could be viewed as deriving entirely from the laryngeal activity (Laver, 1980: 1). This study addresses phonatory modifications, and the term voice quality will thus refer to the laryngeal level only.

The basic type of phonation is modal voice, typical of most speakers; Hollien motivates the term in the following way: "... it includes the range of fundamental frequencies that are normally used in speaking and singing (i.e., the mode)." (1974; cited in Laver, 1980: 109–110). Modal voice is characterised by a neutral phonatory setting; the vibration of vocal folds is periodic without any audible friction and the overall laryngeal tension is moderate (Laver, 1980: 94, 111). This mode of phonation is efficient, with relatively high voice intensity and no special effort required (Skarnitzl, 2016).

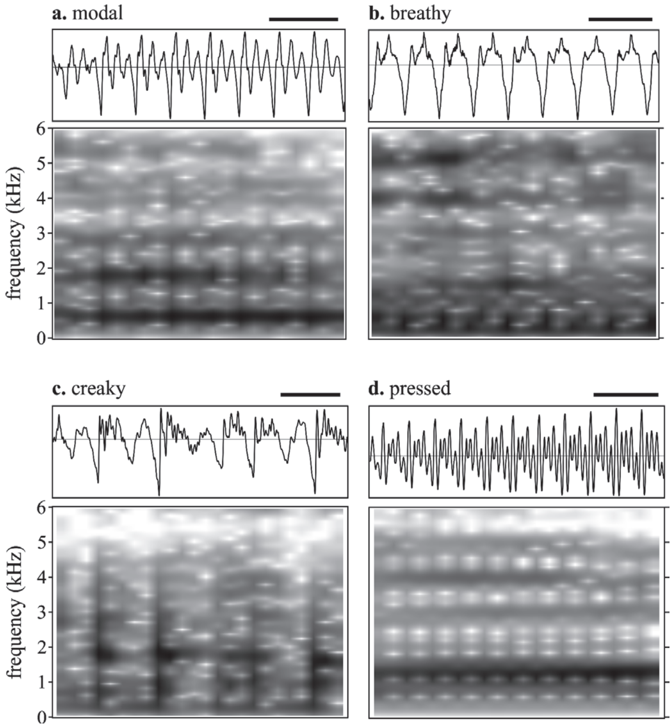
However, the neutral laryngeal setting may be modified both voluntarily and unconsciously reflecting speakers' communication goals (Henton & Bladon, 1985; Anderson, Klofstad, Mayew & Venkatachalam, 2014; Greer & Winters, 2015). Modifications of the neutral phonatory setting may also be caused by changes in speakers' state of health, affective states, or may derive from voice pathology (Tykalová, Rusz, Čmejla, Růžičková & Růžička, 2014). The most common non-modal phonation types differing from the modal one in at least one parameter are breathy voice, creaky voice and pressed voice (Laver, 1980: Chapter 3). These types of phonatory modifications are included in our experiment.

In breathy voice, the mode of vocal fold vibration is inefficient compared with that for modal voice and is accompanied by slight audible friction; vocal folds do not come fully together, which leads to a higher rate of airflow than in modal voice. Consequently, a considerable amount of air is wasted and speakers might need to pause more often to draw breath. Both the intensity and fundamental frequency of breathy voice tend to be rather low (Laver, 1980: chapter 3; Henton & Bladon, 1985). If the laryngeal setting is of permanent nature, it is mostly the case of pathological speech (Shipley & McAfee, 2009; cited in Skarnitzl, 2016). Finally, let us note that women's voices are generally breathier than those of male speakers (Henton & Bladon, 1985; Mendoza et al., 1996), a consequence of differences in the shape of the glottis (Titze, 1989).

Creaky voice represents quite a complex phonation type as there exist several different kinds of it (Keating, Garellek & Kreiman, 2015). Generally, it is characterised by a great irregularity of vocal fold vibrations, low fundamental frequency and intensity, accompanied by creaking and popping noises (Anderson et al., 2014; Abdelli-Beruh, Wolk & Slavin, 2014). Hollien and Wendahl describe the auditory effect of creaky voice as "a train of discrete excitations or pulses produced by the larynx" (1968; cited in Laver, 1980: 124). Catford refers to it as "a rapid series of taps, like a stick being run along a railing" (1964; cited in Henton & Bladon, 1988). Both descriptions imply that the frequency of the vibration typical of creaky voice is so low that listeners can often identify individual pulses.

The last non-modal phonation type to be mentioned here is pressed voice, which involves very high laryngeal tension settings, in some cases even accompanied by hyper-tension of the whole body (Gray & Wise, 1959; cited in Laver, 1980: 129). Pressed phonation is often described as unpleasant, rough, rasping and strident (see studies cit-

Figure 1. Illustration of the four voice qualities described in this section. The horizontal line above each spectrogram corresponds to 10 msec.



ed in Laver, 1980: 127). As in the case of creaky voice, the vibration of vocal folds may be aperiodic, and the voice contains more noise components (Moisik, 2012); however, fundamental frequency (F0) tends to be higher in pressed voice (Laver, 1980: chapter 3). Spectrograms of an open central vowel [aː] pronounced by one of our female speakers in the four voice qualities mentioned above are shown in Figure 1.

1.2 Phonatory modifications in the social context

As mentioned above, various pragmatic reasons may lead speakers to modify neutral phonatory settings while interacting with other people. As for breathy voice, Laver (1980: 135) mentions the paralinguistic use of this phonation type in situations when an interlocutor wishes to communicate messages of a confidential or intimate character. Various studies show that female breathy voices are rated by listeners as more attractive (Henton & Bladon, 1985; Liu a Xu, 2011; Babel, McGuire & King, 2014; Greer & Winters, 2015). Henton and Bladon (1985) argue that British women might want to imitate breathy voice quality in particular communication contexts to increase chances of achieving their goals. Women with a breathy voice may thus be perceived as more desirable and may be given greater recognition by male interlocutors than women speaking with ordinary, modal voice.

The use of creaky voice has become quite widespread in English speakers, especially in the USA (Abdelli-Beruh et al., 2014; Anderson et al., 2014; Greer & Winters, 2015). Greer & Winters (2015) examined the possible social factors behind the increased use of creaky voice by young Americans. They found that creaky quality, traditionally interpreted as a masculine voice quality, contributes to the perception of greater authoritativeness, particularly in young women. Moreover, male speakers with creaky voice are perceived as more "cool" and more attractive. Young Americans may thus exploit this phonation type when attempting to establish authority; additionally, women may tend to use it more to gain the perceived higher status of men.

Some studies show that American women who speak with creaky voice are often very successful and work in the sectors that are traditionally male-dominated, e.g. finance and print media (Carney, 2012; Lepore, 2012, cited in Anderson et al., 2014). Creaky voice, which is characterised by low fundamental frequencies, seems to be exploited when communicating intelligence, seriousness and determination. These findings are similar to those concerning the perception of pitch: speakers with lower-pitched voices tend to be perceived as stronger and more dominant (Puts, Hodges-Simeon, Cárdenas & Gaulin, 2007; Borkowska & Pawlowski, 2011).

Nonetheless, Anderson et al. (2014) conducted an experiment showing a rather negative perception of the creaky voice in American women. 400 male and female listeners from across the United States rated audio recordings of seven women speaking in modal and creaky voice (average age 24). Women using creaky voice were perceived as less competent, less educated, less trustworthy, less attractive, and less hireable, all this regardless of the listeners' gender, age and region.

Finally, pressed voice is often used to signal anger and hostility (Moisik, 2012); a scalar relationship is often suggested between the degree of tenseness and the degree of anger expressed (Laver, 1980: 131–132). It is worth pointing out that some authors talk about harsh voice; this is typically understood as a more extreme setting of pressed voice.

According to Gobl and Ní Chasaide (2003), speakers with pressed voice may also be perceived as stressed, but also confident or even formal. Moisik (2012) argues that harsh voice quality is exploited as a means of representing social identity and stereotypes, namely racial stereotypes of Afro-Americans in the USA.

The survey of literature presented above shows that a speaker's voice quality may be an important tool in his or her self-presentation. On the one hand, the given voice quality may advance a speaker's status, but on the other hand, some voice qualities may affect listeners in a negative way. The aim of this study is to map the effect of the four types of phonation (modal, breathy, creaky and pressed) on the perception and ratings of speakers as job applicants in the job market, using the matched guise technique. Although most speakers use modal voice most of the time, this phonation type can be modified for various reasons. We examine hireability ratings in relation to the different phonation types and personality judgments (likeable, self-confident and trustworthy).

2. Materials and Methods

2.1 Stimuli

Our stimuli were produced by five male (M1–M5) and five female (F1–F5) speakers (average age 25 years, range 19–38 years). The choice of speakers, who were experienced students of phonetics or philology, and phoneticians, was based on their ability to mimic non-modal phonation types. Before recording the stimuli, all the speakers were thoroughly instructed and provided with examples of non-modal voice qualities. They were recorded while reading several repetitions of four different utterances (with an average duration of 15 seconds), each in one of the four different types of phonation (modal, breathy, creaky, pressed). The recordings were made at 48 kHz sampling frequency and 16-bit resolution using an AKG C4500 B-BC condenser microphone in the recording studio of the Institute of Phonetics, Charles University in Prague.

The utterances were designed by the authors so as to sound like answers to questions a job applicant is likely to be asked within a job interview context. Colloquial Czech features were thus used, as illustrated in the following example[1]:

> *Angličtina co se týče takový tý běžný komunikace vůbec není problém. V němčině jsem si jistější, když píšu, než kdybych měl/a třeba s někým mluvit po telefonu. Ale třeba číst maily a odepisovat nebo tak, to je bez problémů; jenom prostě nejsem tak pohotovej/pohotová jako v tý angličtině.*

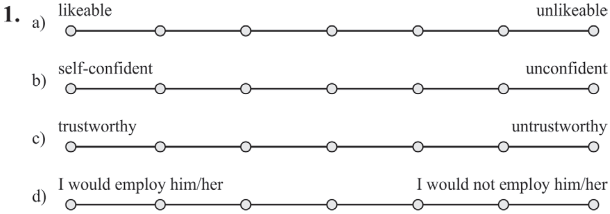
All the recorded stimuli used in the perception test were inspected aurally and visually (using the waveform and spectrogram) by all three authors. The objective of this inspection was to choose each speaker's best rendition of each guise (i.e., phonation type), in other words to ensure the stimuli truly represent the respective voice qualities, as well as to ensure they were free from any speech errors and non-speech noise. An interested reader may find details about some acoustic analyses performed on the selected stimuli – mean F0 and spectral emphasis measured in [a aː] vowels – in the Appendix. The final set of utterances subsequently served as the basis for the perception test in which listeners evaluated the speakers' guises on various dimensions.

2.2 Perception test and participants

The perception test consisted of 40 stimuli (10 speakers x 4 stimuli) which were administered in one of four orders, in four blocks containing 10 stimuli each, with a short pause between the blocks. A short tone was used to signal the onset of each stimulus; the stimuli were then followed by a two-second pause and a desensitization sound.

The listeners were asked to record their ratings of speakers in an answer sheet which contained, for each item, four 7-point semantic differential scales: likeable / unlikeable; self-confident / unconfident; trustworthy / untrustworthy, and I would employ / I would not employ [the speaker], as illustrated in Figure 2. The test itself was preceded by three trial items in which the respondents familiarized themselves with the task. The listeners were instructed that they would hear recordings of various male and female candidates applying for a job position which requires interactions with customers. They were asked to try to rate the speakers based on the sound of their voice rather than the content of the utterances.

Figure 2. A sample item from the perception test (labels translated from Czech).



1. a) likeable ○ ○ ○ ○ ○ ○ ○ unlikeable

b) self-confident ○ ○ ○ ○ ○ ○ ○ unconfident

c) trustworthy ○ ○ ○ ○ ○ ○ ○ untrustworthy

d) I would employ him/her ○ ○ ○ ○ ○ ○ ○ I would not employ him/her

The participants of the perception test were professional employment interviewers and executives from various companies located in Prague who conduct job interviews and make hiring decisions as part of their regular job routine. A total of 38 subjects, 10 men and 28 women (mean age 36.8 years; range 21–57 years), participated in the experiment and were offered 100 Czech crowns as compensation for their participation.

The perception test was administered by the authors of the paper; each participant performed the test individually, in a quiet room using high-quality Sennhesier HD 201 headphones. Praat (Boersma & Weenink, 2015) was used to play the files.

Subsequent statistical analyses and data visualisation were conducted in R (R Core Team, 2016), using the packages *effects* (Fox, 2003) and *ggplot2* (Wickham, 2009).

3. Results and discussion

Overall, it can be stated that the voice manipulations performed by our speakers had a significant effect on the evaluation of the four characteristics, as shown by the results of a repeated measures ANOVA, with Phonation type being the independent variable within the variable Speaker: for likeability, $F(3, 27) = 32.6$; $p < 0.001$; for self-confidence, $F(3, 27) = 48.0$; $p < 0.001$; for trustworthiness, $F(3, 27) = 49.2$; $p < 0.001$; and for employability, $F(3, 27) = 48.1$; $p < 0.001$. More detailed results are illustrated in Figures 3–6 for each personality characteristic.

The results suggest that, perhaps not surprisingly, modal phonation was perceived by our listeners as superior to all other phonation types (in other words, its ratings for likeability, trustworthiness and employability were generally higher; see below for self-confidence ratings). In addition, most of the non-modal guises are evaluated in a very similar way, especially in trustworthiness (Figure 5) and employability (Figure 6).

The most important exceptions to this general finding are visible in the self-confidence ratings (Figure 4). First, eight of the ten speakers were rated similarly for self-confidence in their modal and pressed phonation guise (i.e., modal and pressed phonation scores did not differ significantly); second, breathiness in one's voice impacted self-confidence ratings most negatively. It is interesting to point out that breathy phonation correlates with lower self-confidence ratings not only in male voices but also in female voices. This may be taken as lending indirect support to the study of Anderson et al. (2014) and others cited in section 1.2, according to which creaky phonation – located at the opposite end of the continuum between open and closed glottis configuration than breathy phonation – is associated with confidence and authoritativeness.

Figure 3. Likeability scores for individual speakers in their modal, breathy, creaky, and pressed phonation.

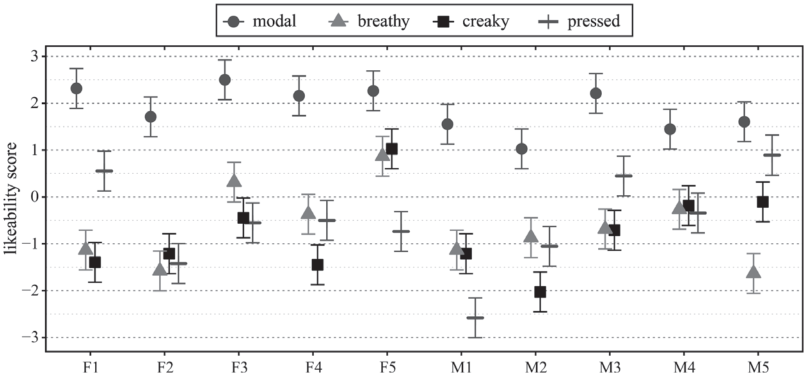


Figure 4. Self-confidence scores for individual speakers in their modal, breathy, creaky, and pressed phonation.

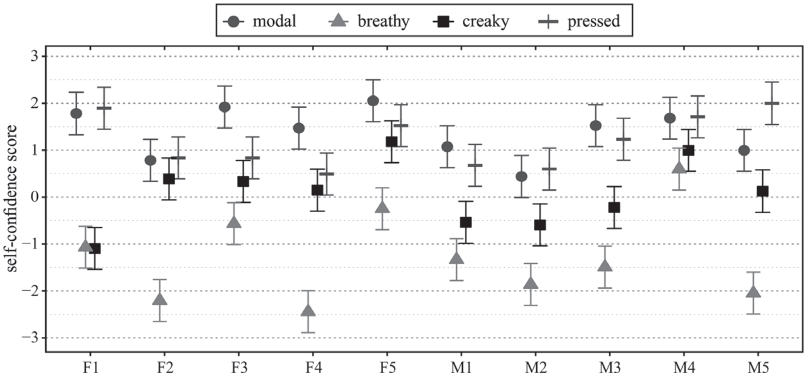


Figure 5. Trustworthiness scores for individual speakers in their modal, breathy, creaky, and pressed phonation.

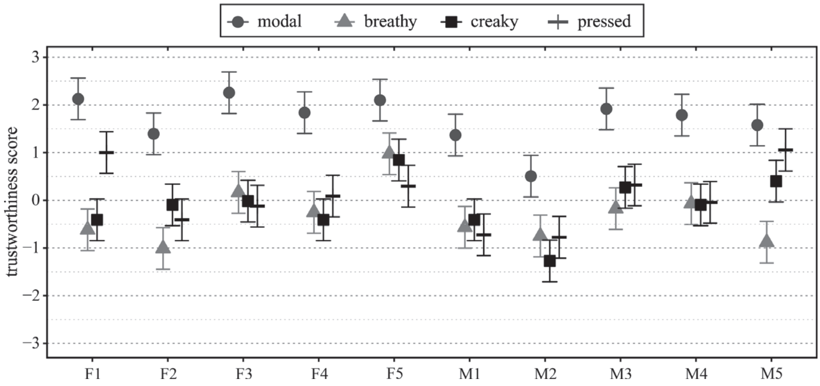


Figure 6. Employability scores for individual speakers in their modal, breathy, creaky, and pressed phonation.

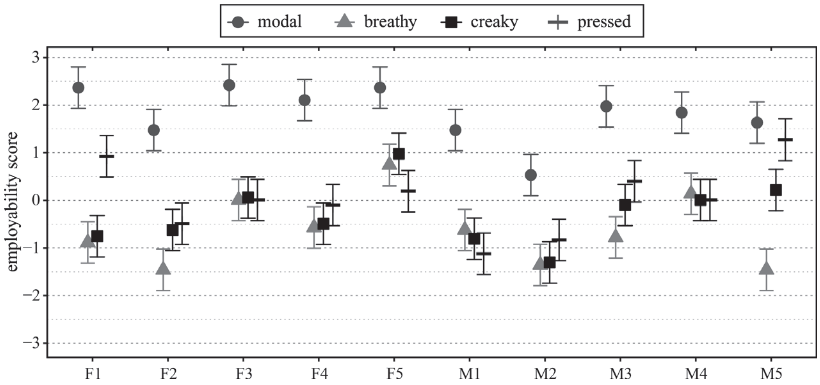


Table 1 shows the summary of posthoc pairwise t-tests, which were conducted for the ratings of the individual dimensions in the four guises. The data confirm what was mentioned above, namely that in most cases the rating of modal phonation significantly differs ($p < 0.05$) from the ratings of the other phonatory modifications, and that pressed phonation is rated differently from the other voice guises on the self-confidence dimension.

Table 1. Summary of posthoc pairwise t-tests, showing which voice qualities were evaluated significantly differently ($p < 0.05$) on which characteristics (L = Likeability, S = Self-confidence, T = Trustworthiness, E = Employability).

FEMALES	modal	breathy	creaky
breathy	L,S,T,E		
creaky	L,T,E		
pressed	L,T,E	S	

MALES	modal	breathy	creaky
breathy	L,S,T,E		
creaky	L,S,T,E		
pressed	T	S	S

Considering the studies mentioned in section 1.2 showing that breathy voice in women is perceived as more attractive than modal voice, we expected this non-modal voice quality to yield higher likeability scores for female speakers. According to our results, however, this is not the case suggesting that likeability and attractiveness do not appear to be simply interchangeable categories. Breathy voice, overall, is not considered as particularly likeable in the job market context; as indicated by informal responses of some of our subjects after the perception test, it rather implies a candidate's low self-confidence. Additionally, it appears to be perceived as projecting submission in men, which is likely to be considered an undesirable personality characteristic. An individual that sounds lacking in confidence or submissive might not be expected to be effective enough when performing his or her job, namely in the customer support branch.

Speakers with pressed phonation guise, on the other hand, were perceived as more self-confident than when using the other non-modal phonation guises, which is in line with Gobl and Ní Chasaide (2003). Given that pressed voice is also associated with anger and hostility, it might be expected to get lower ratings for likeability, and/or possibly for trustworthiness. However, our analysis did not reveal any significant differences in ratings for the two mentioned categories between the non-modal phonation types. Projecting self-confidence may appear to be an important feature in the job market context and could thus affect perceived likeability of the individual's voice.

4. Conclusions

In this study, we focused on voice quality as a means enabling an individual's personality projection and thus having an impact on employers' hiring decisions process. The main result of this study is the contrast in hireability ratings and perceived personal judgments between modal phonation on the one side and the three non-modal phonation types on the other. However, no significant differences were found within the non-modal voices cluster, except for the self-confidence rating.

Future research may thus focus on non-modal phonation types to further explore their effect on the speaker's ratings. It would also be interesting to investigate whether Czech listeners tend to perceive some of these phonation modifications differently from the listeners of different linguistic communities.

ACKNOWLEDGEMENTS

This output was created within the project "Interdisciplinary approach to the linguistic theory issues", subproject "Voice Quality as a Factor of Success in the Job Market" solved at Charles University from the Specific university research in 2016. The third author was supported by the Charles University project Progres 4, *Language in the shiftings of time, space, and culture.*

REFERENCES

Abdelli-Beruh, N. B., Wolk, L. & Slavin, D. (2014). Prevalence of vocal fry in young adult male American English speakers. *Journal of Voice*, 28, 185–190.

Anderson, R. C., Klofstad, C. A., Mayew, W. J. & Venkatachalam, M. (2014). Vocal fry may undermine the success of young women in the labor market. *PloS One*, 9, e97506.

Andrews, M. L. & Schmidt, C. P. (1997). Gender presentation: Perceptual and acoustical analyses of voice. *Journal of Voice*, 11/3, 307–313.

Babel, M., McGuire, G. & King, J. (2014). Towards a more nuanced view of vocal attractiveness. *PloS One*, 9(2), e88616.

Boersma, P. & Weenink, D. (2015). Praat: doing phonetics by computer (Version 5.4.08). Downloaded: May 5 2015, http://www.praat.org.

Borkowska, B. & Pawlowski, B. (2011). Female voice frequency in the context of dominance and attractiveness perception. *Animal Behaviour*, 82, 55–59.

Childers, D. G. & Lee, C. K. (1991). Vocal quality factors: Analysis, synthesis, and perception. *Journal of Acoustical Society of America*, 90(5), 2394–2409.

DeVito, J. A. (2016): *The Interpersonal Communication Book*. Harlow: Pearson.

Fox, J. (2003). Effect displays in R for generalised linear models. *Journal of Statistical Software*, 8(15), 1–27.

Fraccaro, P. J., O'Connor, J. J. M., Re, D. E., Jones, B. J., DeBruine, L. M. & Feinberg, D. R. (2013). Faking it: deliberately altered voice pitch and vocal attractiveness. *Animal Behaviour*, 85, 127–136.

Greer, S. D. F. & Winters, S. J. (2015). The perception of coolness: Differences in evaluating voice quality in male and female speakers. In: *Proceedings of the 18th ICPhS,* Paper 0883.

Gobl, C. & Ní Chasaide, A. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*, 40, 189–212.

Henton, C. G. & Bladon, R. A. W. (1985). Breathiness in normal female speech: Inefficiency versus desirability. *Language & Communication*, 5, 221–227.

Henton, C. G. & Bladon, R. A. W. (1988). Creak as a sociophonetic marker. In: Hyman, L. M. a Li, C. N. (Eds.), *Language, Speech and Mind: Studies in Honour of Victoria A. Fromkin,* 3–29. London: Routledge.

Keating, P., Garellek, M. & Kreiman, J. (2015). Acoustic properties of different kinds of creaky voice. In: *Proceedings of the 18th ICPhS*, Paper 821.

Kreiman, J., Vanlancker-Sidtis, D. & Gerratt, B. L. (2003). Defining and measuring voice quality. In: *Proceedings of VOQUAL'03 ISCA, Tutorial and Research Workshop,* 115–120.

Laver, J. (1980). *The Phonetic Description of Voice Quality.* New York: CUP.

Liu, X. & Xu, Y. (2011). What makes a female voice attractive? In: *Proceedings of 17th ICPhS,* 1274–1277.

Mendoza, E. et al. (1996). Differences in voice quality between men and women: Use of the Long-Term Average Spectrum (LTAS). *Journal of Voice*, 10/1, 59–65.

Moisik, S. R. (2012). Harsh voice quality and its association with blackness in popular American media. *Phonetica*, 69, 193–215.

Puts, D. A., Hodges-Simeon, C., Cárdenas, R. A. & Gaulin, S. J. C. (2007). Men's voices as dominance signals: vocal fundamental and formant frequencies influence dominance attributions among men. *Evolution and Human Behaviour*, 28, 340–344.

R Core Team (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org.

Skarnitzl, R. (2016). Co dokáže náš hlas? Fonetický pohled na variabilitu řečové produkce. *Slovo a smysl*, 26, 95–113.

Skills You Need (2016). *Effective Speaking*. Retrieved on December 22, 2016 from http://www.skillsyouneed.com/ips/effective-speaking.html.

Titze, I. R. (1989). Physiological and acoustic differences between male and female voices. *Journal of Acoustical Society of America*, 85(4), 1699–1707.

Traunmüller, H. & Eriksson, A. (2000). Acoustic effects of variation in vocal effort by men, women, and children. *Journal of the Acoustical Society of America*, 107, 3438–3451.

Tykalová, T., Rusz, J., Čmejla, R., Růžičková, H. & Růžička, E. (2014). Acoustic investigation of stress patterns in Parkinson's disease. *Journal of Voice*, 28(1), 129.e1–129.e8.

Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis (use R!)*. New York: Springer.

NOTES

1. The English version of the provided example of an utterance used in the perception test: *My everyday communication in English, it's not a problem at all. And my German, well, I feel more confident when I write than when speaking to someone on the phone, you know. But, for example, I can read emails and reply to them, that's alright. It's just that I am not as prompt in German as in English.*

APPENDIX

Mean fundamental frequency (F0) and spectral emphasis (SE) computed in Praat from 10 randomly selected [a aː] vowels in the four guises by individual speakers. We used a simplified SE measure: SE = $SPL_{full} – SPL_0$, where SPL_{full} corresponds to the sound pressure level (SPL) of the full spectrum (0–8 kHz) of the given vowel and SPL_0 is the SPL of the low-frequency band cut off at a variable threshold of 1.5 * mean F0 in the vowel (Traunmüller & Eriksson, 2000).

speaker	phonation	F0	SE	speaker	phonation	F0	SE
F1	modal	218.7	-3.9	M1	modal	127.3	-5.9
F1	breathy	244.8	-1.5	M1	breathy	118.8	-1.8
F1	creaky	144.4	-11.3	M1	creaky	93.9	-5.7
F1	pressed	250.7	-7.5	M1	pressed	104.3	-10.9
F2	modal	244.3	-4.2	M2	modal	114.4	-8.2
F2	breathy	259.2	-1.2	M2	breathy	143.3	-5.7
F2	creaky	202.7	-9.6	M2	creaky	95.1	-11.8
F2	pressed	216.7	-8.7	M2	pressed	129.3	-11.1
F3	modal	200.2	-6.4	M3	modal	118.6	-7.7
F3	breathy	184.6	-1.2	M3	breathy	117.5	-3.0
F3	creaky	156.8	-11.0	M3	creaky	105.4	-5.2
F3	pressed	216.3	-9.4	M3	pressed	153.7	-12.0
F4	modal	217.2	-5.8	M4	modal	96.1	-11.4
F4	breathy	201.0	-1.7	M4	breathy	128.8	-6.9
F4	creaky	142.4	-12.0	M4	creaky	82.1	-7.5
F4	pressed	235.1	-7.9	M4	pressed	132.7	-11.1
F5	modal	218.7	-5.7	M5	modal	116.3	-5.7
F5	breathy	190.6	-1.8	M5	breathy	118.9	-1.4
F5	creaky	118.0	-11.7	M5	creaky	99.5	-5.9
F5	pressed	256.4	-8.8	M5	pressed	134.6	-9.3

RESUMÉ

Tato studie přispívá k výzkumu vlivu kvality hlasu na posuzování uchazeče o zaměstnání při přijímacím řízení. Kvalita hlasu je považována za jednu z důležitých složek sebeprezentace na trhu práce, a může tedy šance na uplatnění ovlivnit jak pozitivně, tak negativně. O vlivu různých typů fonace na vnímání mluvčího existuje již řada studií, avšak v českém kontextu se jedná o pole téměř neprobádané. Autoři studie pořídili krátké nahrávky deseti mluvčích (pěti žen a pěti mužů), každého z nich ve čtyřech typech fonací: modální, dyšné, třepené a tlačené. Z těchto čtyřiceti nahrávek byl sestaven percepční test, který byl následně zadán třiceti osmi respondentům, jejichž pracovní náplň zahrnuje účast na přijímacích pohovorech a posuzování uchazečů. Respondenti pomocí sedmistupňové škály zaprvé hodnotili tři osobnostní rysy daného mluvčího (příjemnost, sebejistotu a důvěryhodnost), zadruhé zaznamenávali pravděpodobnost, s jakou by daného mluvčího zaměstnali, přičemž byli instruováni, aby své hodnocení prováděli na základě hlasového projevu mluvčího. Statistická analýza (použita byla korelovaná ANOVA) vliv fonace na hodnocení mluvčího potvrdila; významné rozdíly byly nalezeny především mezi hodnocením modální fonace na straně jedné a zbylých tří typů fonace na straně druhé, a to ve prospěch modální fonace. Hodnocení nemodálních fonací se mezi sebou významně lišila pouze v případě posuzování sebejistoty mluvčího: dyšná fonace se umisťovala na škále sebejistoty nejníže, a tlačená fonace naopak často přibližně na úrovni fonace modální. Možný navazující výzkum by se mohl zaměřit zejména na zmapování případných dalších rozdílů v hodnocení nemodálních typů fonace; či na to, zdali jsou tyto typy fonace v českém kontextu vnímány stejně jako v jiných jazykových komunitách.

Lea Tylečková, Zuzana Prokopová, Radek Skarnitzl
Institute of Phonetics
Faculty of Arts, Charles University
lea.tyleckova@seznam.cz

2017 ACTA UNIVERSITATIS CAROLINAE PAG. 121–137
PHILOLOGICA 3 / PHONETICA PRAGENSIA

SEGMENTAL DURATION AS A CUE TO SYLLABLE BOUNDARIES IN CZECH

PAVEL ŠTURM

ABSTRACT

The aim of the study is to establish whether the acoustic signal contains cues to the syllabification of words that are perceptually relevant, as suggested by previous research. Syllabification preferences of 27 speakers of Czech were examined in a behavioural experiment using disyllabic nonsense words with 10 CC clusters as stimuli. The C1/C2 duration ratio of the intervocalic cluster was manipulated by shortening and lengthening of both consonants. Participants repeated auditorily presented stimuli by syllables, with clear pauses between them (a pause-insertion task). Logistic regression analyses revealed significant effects of sonority type of the cluster, word-edge phonotactics and syllabification strategy reported by the participants in a post-test interview (only half of the participants reported not to have followed any strategy). However, the manipulation condition did not turn out to be a significant predictor, although the C1/C2 ratio correlated negatively with the rate of cluster division. The correlation was in compliance with the hypothesis stating that when C1 is longer than C2, the cluster has a higher probability of being maintained as the onset of the following syllable.

Key words: syllable, syllable boundaries, syllabification, onset maximization, perception, Czech

1. Introduction

Although the phonetic segment, or speech sound, is the smallest recurring linear segment in speech, the processes of human production and perception normally operate at a higher level, namely, the level of words or syllables (Sendlmeier, 1995; Coleman, 2002; Goldinger & Azuma, 2003; Port, 2007). There is mounting evidence that whole stretches of speech are stored as sound-sense associations in memory, which are then recovered in the processing of speech (Coleman, 2002; Goldinger & Azuma, 2003; Hawkins, 2003). This complements the well-known fact that acoustic cues to individual segments are distributed over adjoining segments as well. Phonetic research shows that the syllable plays an important role in speech acquisition, and its language-specific characteristics

https://doi.org/10.14712/24646830.2017.38

also contribute, among others, to the rhythm type of a language. However, determining the location of syllable boundaries presents a considerable challenge to researchers. In phonological description, syllable boundaries are often derived from segments and from theoretical assumptions of the analyst. For instance, the *maximum onset principle* (MOP) predicts intervocalic consonants to belong to the following vowel, thus forming the onset of the syllable, with the provision that phonotactics of the language should not be violated (e.g., Pulgram, 1970; Kahn, 1976). Syllable boundaries can then be effectively viewed as predictable from underlying representations of segments (e.g., Ewen & van der Hulst, 2001: 141ff.).

Experimental evidence is therefore needed regarding the syllabification of words. A growing body of studies suggest that the boundaries of syllables do not appear to follow onset maximization (even in its weaker form) strictly. First of all, phonotactics seems to be gradual rather than categorical, i.e., there are *degrees* of phonotactic legality related to language use and the frequency of occurrence, which is systematically reflected in well-formedness judgments (Vitevitch, Luce, Charles-Luce, & Kemmerer, 1997; Treiman, Kessler, Knewasser, Tincoff, & Bowman, 2000; Munson, 2001; Hay, Pierrehumbert, & Beckman, 2004). Very frequent sequences may be preferred over sequences that occur with relatively low frequency. Secondly, a wide range of both phonetic and phonological factors has been identified in behavioural tasks[1] that affect the syllabification of intervocalic consonants or clusters. We will now review some of these factors briefly.

On the one hand, several phonological effects consistently appear in the experiments. First, intervocalic consonants tend to be attracted to the vowel which is stressed (e.g. Fallows, 1981 and Treiman & Danis, 1988 for English). Second, results from different languages show that phonological length also affects the performance of participants in behavioural tasks: compared to long vowels, short vowels are associated with a higher probability of attracting a coda consonant (e.g. Treiman & Danis, 1988 for English; Schiller, Meyer, & Levelt, 1997 for Dutch; Ní Chiosáin, Welby, & Espesser, 2012 for Irish). Third, assignment to onsets or codas is to some degree influenced by the nature of the intervocalic consonant, specifically by its sonority. In the experiment of Treiman and Danis, sonorant singleton consonants were associated with the previous vowel more closely than obstruent consonants. Similar results were obtained by Goslin and Frauenfelder (2001) for CC clusters in French (obstruent-liquid clusters were treated as complex onsets, whereas other CC clusters were divided) and by Ní Chiosáin et al. (2012) for clusters in Irish (obstruent-liquid *vs*. sonorant-obstruent clusters).

On the other hand, many of these results might be motivated phonetically. For example, the effect of sonority can be related to its acoustic correlates (Parker, 2008; Clements, 2009), or the effect of vowel length to vowel duration. The experiment of Ní Chiosáin et al. (2012) is important because it took into account durational values as well. With increasing duration of the (stressed) vowel in the first syllable, the probability of the vowel

[1] For a detailed discussion of methods, see a review in Côté and Kharlamov (2011). The basic principle of behavioural experiments is that participants perform a simple task that is not directly related to the investigated issue (syllable boundaries). Subjects may double parts of the word, change the order of syllables or repeat one part of the word. For instance, in first syllable reduplication the response /pɪkpɪknɪk/ to the stimulus *picnic* would suggest the syllabification /pɪk.nɪk/, while /pɪpɪknɪk/ would suggest /pɪ.knɪk/. Usually, the term "syllable" is not used, favouring the term "part of word".

attracting a coda consonant decreased. Another study that reports durational effects is that of Redford and Randall (2005). The authors mention another researcher (Christie, 1977) who found that intervocalic clusters were divided when the duration of both consonants in a CC cluster was similar, but kept as an onset when the first C was longer (Redford & Randall 2005: 30). However, the clusters were not word-medial but around word-boundaries (e.g. 'help us nail' [s#n] *vs.* 'help a snail' [#sn]) since Christie focused on juncture cues. In their own experiment using disyllabic nonsense words, Redford and Randall (2005) found that when C1 was shorter than C2, the medial clusters tended to be divided, whereas they were kept together as an onset when C1 was longer. However, this was true only for clusters that were not violating phonotactics – illegal sequences were syllabified invariably according to phonotactics.

The quoted study by Christie is unavailable to us, but Christie (1974) also brings interesting findings. He synthesized 100 tokens of one nonce word ([asta]), varying the formant transients of [a] (flat × more movement), aspiration of [t] (unaspirated × aspirated) and using 25 steps in the duration of the closure interval. The listeners were forced to choose between V.CCV and VC.CV syllabifications. Formant transitions did not seem to have any effect on the listeners. However, the results showed that aspiration of [t] was associated with more C.C syllabifications, which is in accord with the allophonic variation of English plosives where aspiration after [s] in the same syllable is disallowed. Moreover, there was a gradual increase in the proportion of C.C syllabifications in response to increasing duration of the closure interval of [t], i.e., along with lengthening the second consonant of the cluster. This suggests that the C1/C2 ratio is indeed important in syllabification judgments (at least in synthetic speech).

The aim of the current experiment is to replicate such findings with Czech listeners using acoustic manipulations of the signal. Redford and Randall (2005) did the durational analysis ex post, taking the variability in duration into account. Our experiment is instead designed to investigate this effect explicitly. In addition, the authors used a written task, in which syllable boundaries were marked by the subjects on paper (the listeners were asked to write down nonce words that they heard and divide them into syllables). This is burdened with metalinguistic awareness to a higher degree than the behavioural experiments reported above (see Goslin & Frauenfelder, 2001). In contrast to Christie (1974), we would like to investigate a wider range of clusters, which also increases ecological validity of the experiment.

It is reasonable to believe that a significant array of phonetic features in the acoustic signal may contribute to the perception of boundaries between syllables. However, we focus only on the C1/C2 durational ratio. H0 would state that syllabification of medial clusters is not influenced by the temporal structure of the cluster. In contrast, our alternative hypothesis (H1) predicts that there will be more V.CCV syllabifications for tokens where the ratio has been raised (or less V.CCV syllabifications where it has been lowered). In other words, the longer the C1 (or the shorter the C2), the more CC onsets are predicted compared to unaltered tokens. This assumption is substantiated by the literature reported above and in general by domain strengthening effects, where for instance segments close to the initial boundary are longer or differently articulated (Fougeron & Keating, 1997).

In addition to this main line of interest, some of the other factors shown to be affecting syllabification judgments can be examined. Given the practical limitations of the experiment in terms of time and scope, we do not investigate the effects of stress or the phonological length of the preceding vowel. However, the target clusters vary in their frequency of occurrence word-initially from very frequent clusters to clusters that do not occur at all, which allows us to evaluate the contribution of word-edge phonotactics. It is predicted that sequences with a higher frequency of occurrence will be more likely to be preserved as CC onsets than less frequent sequences (Hypothesis 2). Moreover, the clusters are of different sonority and manner of articulation types (combinations of stops, fricatives and sonorants). Therefore, the sonority sequencing principle can be taken into account as well. Hypothesis 3 thus states that clusters with a rising sonority (obstruent-sonorant sequences) will be more likely to be preserved as CC onsets than clusters with a plateau in sonority (sequences of two obstruents or two sonorants).

2. Method

The material comprised a set of nonsense words that were modelled on the Czech language. Most of the pseudo-words differed from genuine words by a single phoneme substitution (e.g., /zolba/ ~ *rolba*), but several items included more substitutions (e.g., /knɛtrɛm/ ~ *svetrem*). The test session included 10 target nonsense words that represented 10 unique clusters differing in frequency and manner of articulation:

- two items with a plosive-plosive sequence (/t͡ʃaktɛm/, /natka/)
- one item with an affricate-plosive sequence (/smat͡skɪ/)
- three items with a fricative-plosive sequence (/lɛsta/, /vɪskɛm/, /zaxtɪ/)
- two items with a plosive-sonorant sequence (/knɛtrɛm/, /xɛbra/)
- one item with a fricative-sonorant sequence (/kɛslo/);
- one item with two sonorants (/xarmu/).

A set of distractor items varying in the number of syllables and cluster length was added to the test session as a source of variability (/lo͡uʃc/, /fnus/, /mostlɪna/, /kotxarma/, /toʒɲɪt͡sɛ/, /xɛt͡ʃɛ/, /saːnɛf/); these items assisted to conceal the medial CC sequences in target items. In addition, a set of training items was constructed that was used for the initial pre-test session in which the participants familiarized themselves with the range of material to be presented later (/vraːs/, /vufɛ/, /t͡ʃɛka/, /zolba/, /mɛspɛkt/, /kɛjsɛk/, /dorsko/, /loskamɛ/, /paːnɛsto͡uʃ/, /nɛspozɪn/).

A female native speaker of Czech (22 years old) read a list of nonsense words including distractor items (hereafter we will refer to them simply as “words”). The experimenter was present during the recording, which took place at the Institute of Phonetics in Prague (sound-treated recording booth, condenser microphone, 16-bit 32-kHz audio). When a correction was necessary, the experimenter asked the speaker for a new version of the item. The aim was that the final speech production should sound as natural as possible, without emphasis, without syllable lengthening.

The acoustic signal of all target items and of selected other items was manipulated in *Praat* (Boersma & Weenink, 2014). First, the boundaries of the two intervocalic consonants

(C1 and C2) and the preceding vowel (V) were determined, and the duration of these segments was extracted (see Appendix). The rules for segmentation of the acoustic signal can be summarized as follows (see also the recommendations in Machač & Skarnitzl, 2009):

- the decisive factor was the formant structure, which defines vowels and sonorants; the boundary between an obstruent and sonorant segments was therefore placed at the point where full formant structure began/ceased to appear (in ambiguous cases, the boundary was placed in the middle of the transitory region);
- nasals have full formant structure, but are associated among others with a drop of energy in the higher frequencies and the presence of a nasal formant;
- lateral approximants are very similar to vowels, but are associated with lower formant values and a drop of energy in the higher frequencies; when the lateral was visually indistinguishable from the vowel, the decision was based primarily on audition alone; however, the sequence [sl] is problematic because of the synchronization of articulatory gestures – we can distinguish friction of [s], friction of devoiced [l̥] and full formant structure of [l]; the lateral in our analysis included both the [l̥] and [l] parts.

Figure 1. A spectrogram of the stimulus [kɛslo] with marked boundaries of target speech sounds. The devoiced ([l̥]) and sonorant ([l]) parts of the liquid are considered to be one speech sound.

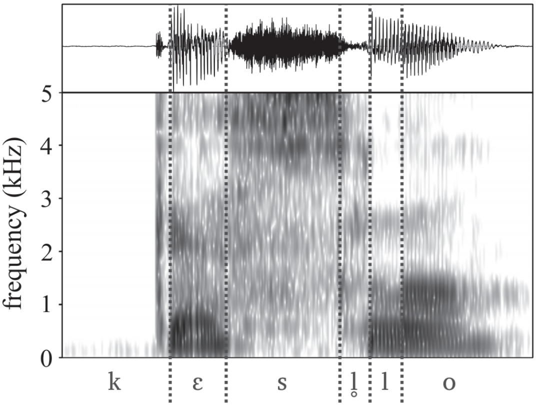


Figure 1 shows the segmentation of [kɛslo] as an example of boundary placement in plosive-vowel, vowel-fricative, [sl] and lateral-vowel contexts.

In the next step, five Manipulation objects were created from the Sound object (using default parameters): with *original* C1 and C2 durations (relative duration = 1.0), with C1 or C2 *lengthened* by half of its duration (relative duration = 1.5), and with C1 or C2 *shortened* by half of its duration (relative duration = 0.5). The duration of the preceding vowel was not altered. PSOLA resynthesis was used to create new audio files from these Manipulation objects. The perceptual test thus included 18 training items, 26 distractor items and 50 target items (5 manipulations × 10 words), totalling 94 items.

The perceptual test was administered in DMDX (Forster & Forster, 2003) without any information displayed on the screen. Each new item was introduced by a short warning signal (a combination of noise and tones), which also functioned as a simple means of perceptual desensitation. After 800 milliseconds the participants heard the stimulus itself, and their response was recorded (see below for the instructions). The time to respond was limited to 3–4 seconds. Individual items were played automatically so that no further activity was necessary on the side of the participants. The experiment was divided into a training session followed by four test blocks. The participants were prompted to take a short break after each block. The order of the 19 items in a block (and the order of the blocks themselves) was randomized for each participant. The total duration of the experiment did not exceed 17 minutes. In a post-test interview, the participants were asked to report whether they had followed some strategy in the task.

The task was to *repeat the presented word by syllables*, with clear pauses between them. The participants were asked to follow only their first impression of the sound, how they perceived it. They were urged to listen very carefully to the stimuli. The participants were told that they were going to hear different variants of individual words, which were not supposed to have the same characteristics (or outcome of the division into syllables). They should consider and divide each word separately, individually, without reference to previous cases. The speech production of the participants was recorded with a microphone, and the location of syllable boundaries was identified with the aid of this recording (with silence between sound intervals implying a syllable boundary percept on part of the listener).

27 subjects participated in the experiment (19 females, 6 males, median age = 21 years), all students of English at a pedagogical faculty. However, two subjects were discarded prior to performing any analyses, one for being bilingual and one for showing signs of miscomprehension of the task during the training session. Thus, only 25 speakers were analysed (yielding 1250 tokens). Furthermore, 63 tokens (5% of the data set) with missing or ambiguous syllabification were removed. This comprised cases where, for instance, the speaker produced a given word without a break between the syllables, or hesitated, or produced a different consonant in the target cluster. The final number of analysed tokens was therefore 1187.

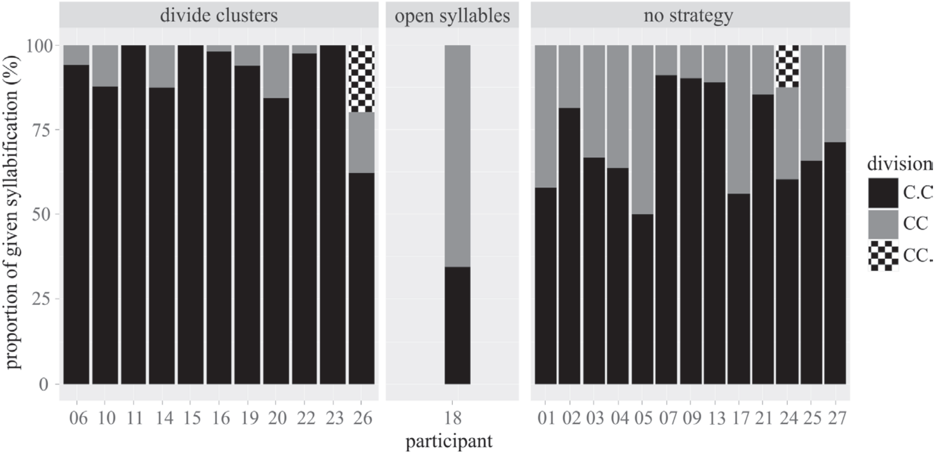
Statistical analyses were performed in the *R* software (R Core Team, 2016) using the package *lme4* (Bates, Mächler, Bolker, & Walker, 2015). Figures were drawn using the package *ggplot2* (Wickham, 2009). The data were analysed with a logistic mixed-effects regression that allows investigating the effect of predictors on a binary dependent variable (the type of syllabification outcome: V.CCV × VC.CV). Individual predictors are introduced in the results section. The statistical significance of a predictor was evaluated in a goodness-of-fit test using maximum likelihood ratio by comparing the full model (a given set of predictors and their interactions) with a reduced model (lacking one predictor or interaction). Maximum random-effect structure was used that still allowed the model to converge. In addition, other basic statistical functions were used in the analyses (t-tests, correlations, binomial tests).

3. Results

Overall, the preference was to divide the intervocalic clusters between two syllables (79%), followed by V.CC syllabifications (20%) and CC.V syllabifications (1%, $n = 15$). Given that three speakers divided the intervocalic clusters in all stimuli (thus yielding

100% of C.C syllabifications), it is likely that individual participants may have different strategies. Figure 2 therefore shows the response patterns of all participants depending on the strategies reported in the post-test questionnaire. Approximately half of the participants said they did not follow any specific strategy in the task. A similar number of participants admitted to divide any intervocalic clusters. Only one participant reported that he endeavoured to pronounce open syllables, keeping the cluster as an onset. Accordingly, this speaker is associated with the lowest proportion of C.C responses, and the "no strategy" group generally seems to yield a lower proportion of C.C responses than the "divide clusters" group. Interestingly, the CC.V responses were produced by only two speakers. Given its speaker specificity and low occurrence, this category was therefore also excluded from the results, leaving 1172 tokens for analysis using logistic regression with a binary response variable. (However, we will return to the VCC.V syllabifications in the Discussion.)

Figure 2. Proportion of syllabification response types for individual speakers in relation to their reported strategies in the experimental task.



The logistic regression analysis therefore includes strategy as a predictor, which proved to be statistically highly significant ($\chi^2(2) = 21.9$, $p < 0.001$). Adding this predictor reduced the variance of the random effect of participant (from 2.7 to 0.9). The goodness-of-fit of the model was further improved when sonority was included as a three-level predictor ($\chi^2(2) = 16.2$, $p < 0.001$), differentiating between clusters of two obstruents × two sonorants × an obstruent followed by a sonorant. The residual variance of the random effect of word decreased from 2.0 to 0.3. Specifically, S-S clusters were associated with the highest odds of division, whereas O-S clusters with lowest. The sonority effect was also added as a slope to participant, allowing individual participants to differ in sensitivity to the sonority classes ($\chi^2(5) = 15.1$, $p < 0.01$). Further, it is likely that the frequency of occurrence of the cluster may play a role in the syllabification behaviour of the participants. The predictor of frequency – log ipm frequency of occurrence of the sequence as a word-initial onset, adopted from Šturm and Lukeš (2017) – was therefore added to the model, which increased its goodness-of-fit significantly ($\chi^2(1) = 4.6$, $p < 0.05$). However, the direction of the influence was unexpected: more frequent clusters had somewhat higher odds of division than less frequent clusters. There was no significant interaction of FREQUENCY and SONORITY. Figure 3 shows the effect plots for STRATEGY, SONORITY and FREQUENCY in terms of the probabilities of cluster division (= C.C syllabification).

The main investigated factor was MANIPULATION – we predicted that syllabification would be affected by changes in the temporal relation between C1 and C2 in the intervocalic cluster. The overall results do not support this conclusion: adding this effect into the model did not increase its goodness-of-fit significantly. Moreover, the interaction term for MANIPULATION*SONORITY was not significant either. However, the manipulation of C1 and C2 duration cannot be treated as equal for all items, since lengthening or shortening may not always change the value for the C1/C2 ratio. Therefore, we substituted MANIPULATION with a binary parameter ASYMMETRY (C1 is longer × C1 is not longer)

Figure 3. Probability of cluster division (C.C syllabification) as a function of strategy (left), sonority type (middle) and log frequency (right). (O = obstruent, S = sonorant)

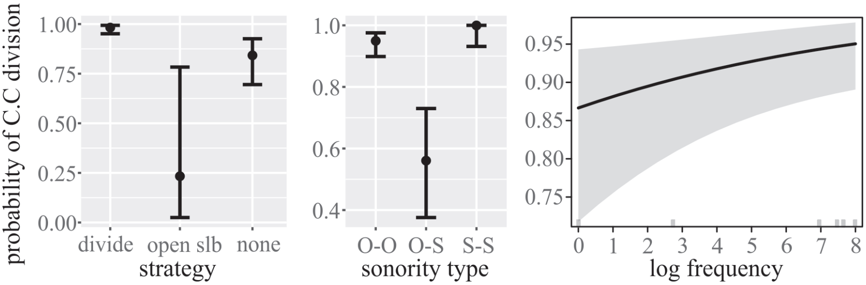


Figure 4. Correlation between C1/C2 ratio and proportion of cluster division (C.C syllabification) computed on word-by-manipulation cells. The regression line is indicated along with 95% confidence intervals of the regression model. Only speakers without a reported task strategy ($n = 13$).

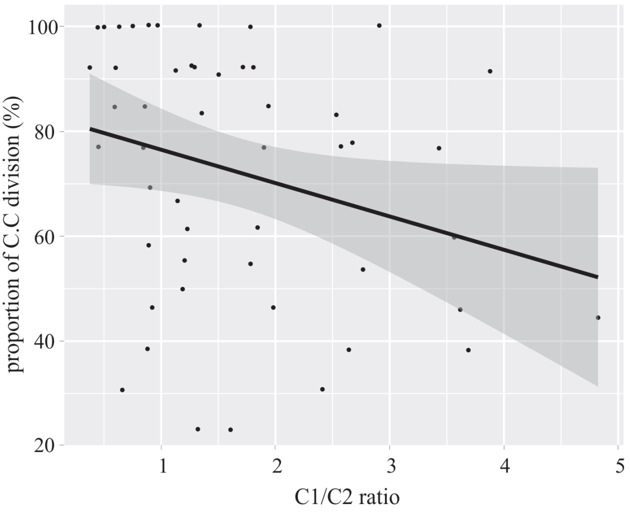
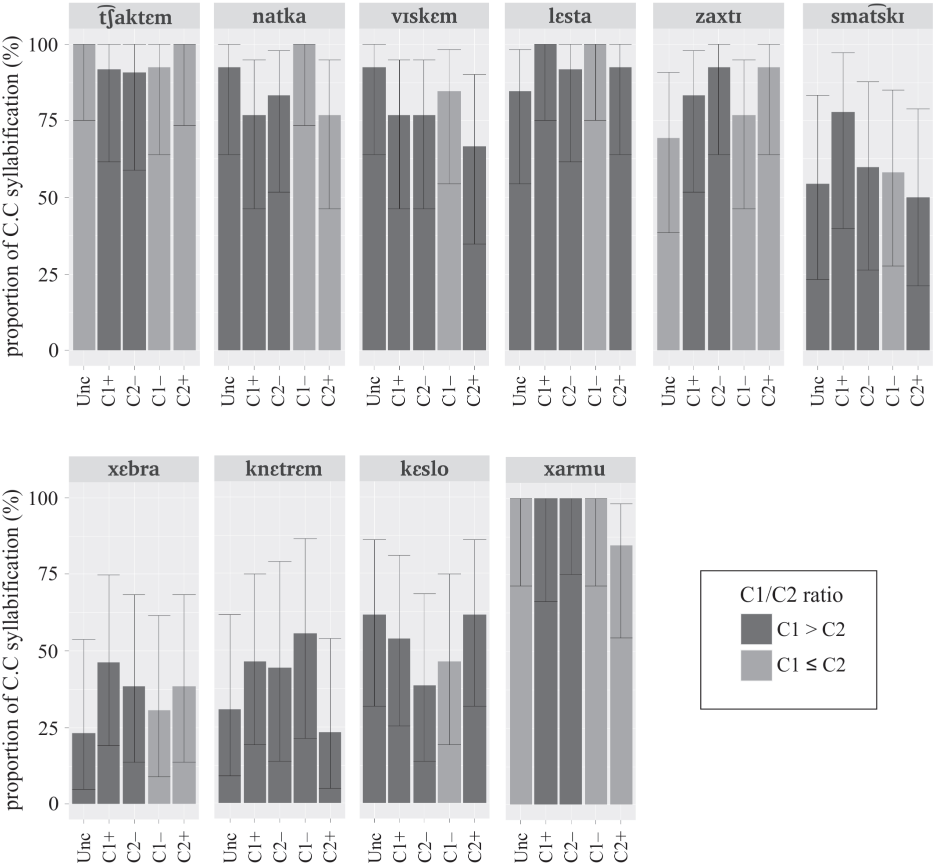


Figure 5. Proportion of cluster division (C.C syllabification) as a function of individual words and manipulation condition (Unc = unchanged, C1/C2 = 1st/2nd consonant, +/– = lengthened/shortened). The colour indicates whether or not the resulting C1/C2 ratio was greater than one (C1 > C2). The whiskers indicate 95% confidence intervals from a binomial test. Only speakers without a reported task strategy (*n* = 13).



in order to see whether the acoustic structure of the manipulated clusters is relevant. Although it reached smaller *p*-values than manipulation, it was far from statistical significance ($\chi^2(1) = 0.2$, $p = 0.66$). Lastly, we filtered out all participants that reported to follow some strategy, narrowing the analysed sample to data from 13 participants. The effects observed in the subsample were by and large identical to the previous results, confirming the effects of sonority and frequency, and confirming the lack of effect of asymmetry. Thus, not even the speakers without a task-related strategy seemed to be influenced by the acoustic manipulations in their syllabification behaviour.

A correlation analysis based on the subsample nevertheless showed a statistically significant relationship between the C1/C2 ratio and the proportion of C.C responses ($r = -0.28$, $p < 0.05$). This is visually represented in Figure 4: the higher the C1/C2 ratio, the lower the rate of cluster division (i.e., a greater preference for CC syllable onsets). However, the weak correlation coefficient indicates that only 8% of the variance in the parameters could be explained. Additionally, the ratio was transformed into a binary variable *asymmetry* (like in the logistic model); listeners divided the stimuli with C1 longer than C2 68% of the time, while it was 80% of the time for stimuli with C1 shorter than or equal to C2 (the difference was not significant in a t-test, $p = 0.12$). The general preference, notwithstanding the C1/C2 ratio, is thus for cluster division. Durational manipulations of the items seem to exert only a small influence on the participants.

Finally, Figure 5 shows the proportion of cluster division for individual words and manipulations. The colour of the bars indicates whether or not the given item has C1 longer than C2 (for instance, the word [knɛtrɛm] had consistently longer C1s in all conditions, i.e., even shortening of C1 did not change the direction of the C1/C2 ratio). It is immediately apparent that the confidence intervals completely overlap for the manipulations within words, suggesting no change in syllabification across conditions. The only trend is that O-S clusters seem to behave differently from O-O or S-S clusters. With a possible exception of the word [smat͡skɪ], individual words in a sonority group do not diverge substantially. Despite the lack of clear evidence suggesting an effect of manipulation in the hypothesized direction (higher C1/C2 ratio will lead to lower rate of cluster division), an important finding is that there was simultaneously no opposite effect, i.e. a lower ratio being associated with lower rates of division.

4. Discussion

The aim of the experiment was to establish whether the acoustic signal contains cues to the syllabification of words that are perceptually relevant. This has already been approached in previous research (Christie, 1974; Redford & Randall, 2005; also Christie, 1977 cited in Redford & Randall, 2005), but has not been investigated for Czech. The latter two studies concluded that when the first member of a two-consonant cluster is longer than the second member, listeners have a tendency to treat the cluster as the onset of the following syllable; in the opposite case, the probability of division of the cluster increases. Similarly, Christie (1974) discovered, on a synthesized speech token, that increasing the duration of the closure interval in intervocalic [st] was associated with a gradual increase in the proportion of C.C syllabifications.

The current experiment was designed to replicate these findings with Czech listeners. We repeat the hypotheses of the study here for the sake of convenience:

- H1: more V.CCV syllabifications are expected for tokens where the C1/C2 ratio is raised (or less V.CCV syllabifications where it is lowered);
- H2: sequences with a higher frequency of occurrence are more likely to be preserved as CC onsets than less frequent sequences;
- H3: clusters with a rising sonority are more likely to be preserved as CC onsets than clusters with a plateau in sonority.

The first hypothesis was not confirmed in the statistical analysis based on the whole data set. Manipulated items did not differ significantly from non-manipulated items. Depending on the duration ratios in the original stimulus, a critical boundary (C1/C2 = 1) could be crossed by the manipulation, but it was not the case in all the words. Therefore, the lack of a manipulation effect could be explained in some items by the stability of the ratio. However, not even categorizing individual tokens into "C1 longer than C2" *vs.* "C1 not longer than C2" suggested any significant changes in the response patterns.

Furthermore, the extent of the acoustic manipulations was quite massive and above the just noticeable difference. We can expect that in natural speech the differences in duration would be of lower magnitude, which would obscure the potential syllable boundary effect even more. This suggests that the null hypothesis should not be rejected (H0: syllabification is not influenced by the duration of consonants in an intervocalic cluster). In a similar vein, the authors mentioned above admit that the results only concerned words for which the syllabification was ambiguous (Redford & Randall, 2005: 42–43), i.e., when there were several syllabifications options, all of them allowed by the phonotactics of the language. For the most part, our sample included precisely these cases (even the words /xarmu/ or /t͡ʃaktɛm/ could be syllabified in other ways than C.C: compare the initial onsets in *rmoutit* /rmo͡ucɪt/ or *který* /ktɛriː/). We can expect that syllabification of illegal sequences, ruled out by the phonotactics, would be even more resistant to acoustic manipulations, favouring invariably the C.C division. However, this prediction seemed to be compatible only with the illegal cluster in /zaxtɪ/ but not with the illegal cluster in /smat͡skɪ/, which was, quite unexpectedly, associated with a significant amount of V.CCV responses (/t͡sk/ will be discussed in detail below).

Importantly, it must be stressed that we only focused on the duration of consonants in the cluster. Thus the manipulations involved stretching or shortening of C1 (or C2) duration by 50%, while no manipulations were performed on the vowels. Yet it is clear that perception utilizes many other cues apart from the temporal structure of the intervocalic cluster, e.g. the C1/V1 ratio (see Kingston, Kawahara, Chambless, Mash, & Brenner-Alsop, 2009 for geminates; Maddieson, 1985). However, in a preliminary analysis this parameter did not seem to contribute to the syllabification responses in any way.

A possible reason behind the lack of a manipulation effect could be that some participants reported to have followed a certain strategy in the task, which might have reduced their sensitivity to the acoustic manipulations. Although the effects observed in a subsample of 13 participants – those who reported to have "no strategy" – were by and large identical to the previous results, a correlation analysis showed a statistically significant relationship between the C1/C2 ratio and the proportion of C.C responses (specifically, there was a greater preference for CC onsets with higher C1/C2 ratios). This finding at least is in accord with H1. Moreover, the results of our experiment – both of the subsample and of the whole data set – do not contradict the H1 in the sense of opposite direction. Although durational manipulations of the items seemed to exert none or only a small influence on the participants, there was simultaneously no change in syllabification across conditions that would suggest that a *lower* C1/C2 ratio is associated with *higher* rates of V.CCV syllabifications.

The prediction of H2 was not borne out. On the contrary, a positive – not negative – correlation was ascertained between cluster frequency of occurrence and the probability of cluster division. For instance, the relatively frequent clusters /st/ and /sk/ were most often split into two syllables, which counters the expectation. Moreover, the phonotactically illegal cluster /t͡sk/ was not predominantly split, as would be expected, but was ambivalent between C.C and V.CC syllabifications. Thus, a substantial number of participants[2] produced an ill-formed onset cluster in response to the word /smat͡skɪ/, which seems to contradict the well-formedness principle whereby only syllables encountered at the edges of a word result from word-medial syllabification. If this is the case, then either the phonotactic principle should not be given such a prominent place in syllabification, or speakers might not perceive the /t͡sk/ cluster as illegal (e.g., they might treat its absence from Czech word onsets as an accidental gap). Figure 5 reveals that the patterns of /t͡sk/ were quite similar to the obstruent-sonorant clusters, especially /sl/, but the data offer no clear explanation for this behaviour, apart from the fact that /t͡sk/ is the only cluster with an affricate sound as its member. However, Šturm (2017, p. 89) found in a similar experiment using genuine Czech words that the proportion of C.C syllabification of the clusters /t͡sk/ and /t͡ʃk/ was higher, approximately 80%, which is more in line with the hypothesis. Therefore, an alternative explanation concerns the material used: since the participants in the current experiment responded to nonsense words, they might have treated the sequences differently from real speech material (namely, with more benevolence towards certain sequences). This will be discussed in more detail below.

With regard specifically to the cluster /kt/, one explanation may relate to how the frequency was computed. The counts are based on written corpora, where the cluster /kt/ – associated with one of the highest rates of C.C division in the experiment – is more common than in spoken corpora due to the frequent use in written texts of relative clauses with the pronoun *který* ("which/who"). However, it must be admitted that a separate experiment is needed for investigating the effect of cluster frequency on syllable division. In the current state, only 10 clusters were taken into account, which is too small a number given that there were also differences in sonority and manner of articulation that might represent more decisive factors in syllabification.

H3, concerning the sonority type of the cluster, was substantiated by the results of the experiment and is completely in compliance with previous findings (Goslin & Frauenfelder, 2001; Ní Chiosáin et al., 2012). Obstruent-sonorant clusters were most frequently maintained as onsets, whereas clusters of two obstruents were more often divided, and the cluster of two sonorants was divided almost always. In fact, the difference between the latter two – i.e. clusters with sonority plateaus – was not substantial. The only exception was /t͡sk/, which has already been discussed. It is especially interesting to compare plosive-plosive clusters with fricative-plosive clusters. In several approaches to the sonority hierarchy, but not in some others, fricatives are placed higher on the scale than plosives (Zec, 2007: 178; Gordon, 2016: 99). Since the /sk st xt/ clusters would then violate the sonority profile in the syllable onset, we may expect a higher proportion of C.C syllabifications compared to the plosive-plosive clusters. This seems to be the case possibly with /st/, but not with the other two clusters. However, the very similar behaviour of both types of clusters does not necessarily present a case for treating fricatives and plosives together as obstruents because sonority plateaus, represented by the plosive-plosive sequences, are avoided as well. In other words, both approaches to sonority classes would lead to the same conclusion, namely, a strong preference of C.C. syllabifications.

[2] Eight out of thirteen participants in the subset. Interestingly, all but one were associated with the highest proportions of CC onsets generally according to Figure 2. Therefore, the increased number of CC syllabifications might reflect their general strategy rather than the specifics of the /t͡sk/ cluster.

However, the case of /sl/ is different. The concept of minimal sonority distance (see Zec, 2007) assumes that a certain difference in sonority is needed between the first and second member of a CC onset. If fricatives are higher on the sonority scale than plosives, then the difference from liquids is smaller for /s/ than for plosives, implying that e.g. a /pl/ cluster is more ideal than a /sl/ cluster. Although Figure 5 suggests that /sl/ might be more often divided by the participants than /br/ and /tr/, the difference is not significant. Moreover, it could be the case that the ambiguous syllabification of /kɛslo/ (bordering around 50% of C.C or CC responses) is related to the ambiguous phonetic segmentation of the cluster, as shown in Figure 1. We do not know whether the devoiced part of the lateral belongs, perceptually, to the fricative or to the approximant. An analysis of individual participants revealed that the V.CCV syllabification was again especially linked to those speakers who, compared to the other speakers in the subset, generally produced more CC onsets (Fig. 2, see also note 2). With regard to the extreme cases, speakers 5 and 17 syllabified all five tokens of /kɛslo/ as V.CCV, whereas speakers 7 and 13 syllabified them in all instances as VC.CV. Five other speakers showed a less strong preference for one of the options, and four speakers did not incline to either of the options.

With the exception of the /rm/ cluster, which was almost unanimously divided, the syllabification outputs were not clear-cut (dichotomous, either-or). Since eight out of the ten clusters were phonotactically legal (including /rm/, although it is not frequent word-initially), the prevailing strategy should have been that of *onset maximization*. This is a common assumption of many researchers and writers about syllabification (Pulgram 1970; Kahn, 1976; Fallows, 1981; Hall, 2006; see also Bičan, 2017), but it is clearly contradicted by the results of the current experiment and of other experiments in Šturm (2017). The clusters were syllabified – regardless of whether the two illegal clusters were included – as VC.CV in 78% of cases, whereas the V.CCV division, conforming to onset maximization, occurred in only 20% of the cases. Moreover, there were 15 cases (1%) of VCC.V syllabification. Although we might discard these outputs as marginal (and we indeed excluded them from the main analysis), the syllabification pattern nevertheless occurred and it represents further evidence against the maximal onset principle. The data come from two participants (S24 and S26) and 6 words/clusters (in descending order of frequency of occurrence: /xarmu/, /zaxtɪ/, /t͡ʃaktɛm/, /vɪskɛm/, /lɛsta/, /natka/). Four of the clusters display a falling sonority pattern typical of syllable codas, and two have sonority plateaus. Thus, despite the speaker-specificity, the occurrence cannot be declared unexpected or unnatural.

Our decision to use nonsense words entails the acceptance of some assumptions along with it. As pointed out by a reviewer, one implicit assumption is that Czech participants pronounce non-Czech words like Czech words. Yet it is common to experience difficulties – and change the tempo or manner of speech, for instance – when we encounter an unfamiliar word in a text. This may have contributed to the significant number of invalid responses in the experiment, such as hesitations or slips of the tongue. However, the participants were expecting nonsense words to appear because they familiarized with the task and the range of stimuli in the training session. Moreover, there is no reason to believe that the participants would pronounce a nonsense word like /xarmu/ any differently from the Czech word /ʃarmu/. The second assumption, saying that Czech participants *syllabify* non-Czech words like Czech words, is more difficult to substantiate. We can only highlight again the close similarity between the nonsense words used and genuine Czech words. Also, the results of the experiment in terms of the sonority factor seem to suggest that the two types of stimuli are to a large degree treated similarly (but note the deviant cluster /t͡sk/). A crucial difference is the exclusion of lexical information; however, this is rather beneficial, since it prevents any morphological effects from influencing the results.

Finally, 12 out of 25 analysed participants reported that they followed a strategy in the behavioural task, such as onset maximization or cluster division, despite being explicitly instructed not to do so. There were no significant differences between male and female participants, except that males had a greater tendency towards CC onsets regardless of the stated strategy. Moreover, it is questionable whether the 13 remaining participants really performed the task "without a strategy", as they stated. This might be a serious limitation to the study. Figure 2 suggests that the "no strategy" group forms two clusters of listeners, with five participants resembling the performance of subjects from the "divide clusters" strategy. However, it is very difficult to persuade participants to break free from all syllabification rules and other deep-rooted habits, if not outright impossible. In any case, future research would benefit from a larger sample of participants.

ACKNOWLEDGEMENT

This research was supported by the Charles University project Progres 4, *Language in the shiftings of time, space, and culture*.

REFERENCES

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bičan, A. (2017). Slabikování. In P. Karlík, M. Nekula, & J. Pleskalová (Eds.), *Nový encyklopedický slovník češtiny online*. Retrieved from http://www.czechency.org

Boersma, P., & Weenink, D. (2014). Praat – doing phonetics by computer (Version 5.4). Retrieved from http://www.praat.org

Christie, W. M. (1974). Some cues for syllable juncture perception in English. *Journal of the Acoustical Society of America*, *55*(4), 819–821.

Clements, G. N. (2009). Does sonority have a phonetic basis? Comments on the chapter by Bert Vaux. In E. Raimy & C. E. Cairns (Eds.), *Contemporary Views on Architecture and Representations in Phonological Theory* (pp. 165–175). Cambridge, Mass: MIT Press.

Coleman, J. (2002). Phonetic representations in the mental lexicon. In J. Durand & B. Laks (Eds.), *Phonetics, Phonology, and Cognition* (pp. 96–130). Oxford: Oxford University Press.

Côté, M.-H., & Kharlamov, V. (2011). The impact of experimental tasks on syllabification judgments: A case study of Russian. In C. Cairns & E. Raimy (Eds.), *Handbook of the Syllable* (pp. 273–294). Leiden: Brill.

Ewen, C. J., & van der Hulst, H. (2001). *The Phonological Structure of Words: An Introduction* (Vol. 39). Cambridge: Cambridge University Press.

Fallows, D. (1981). Experimental evidence for English syllabification and syllable structure. *Journal of Linguistics*, *17*(2), 309–317. https://doi.org/10.1017/S0022226700007027

Forster, K. I., & Forster, J. C. (2003). DMDX: a windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers: A Journal of the Psychonomic Society, Inc*, *35*(1), 116–124.

Fougeron, C., & Keating, P. A. (1997). Articulatory strengthening at edges of prosodic domains. *The Journal of the Acoustical Society of America*, *101*(6), 3728–3740. https://doi.org/10.1121/1.418332

Goldinger, S. D., & Azuma, T. (2003). Puzzle-solving science: The quixotic quest for units in speech perception. *Journal of Phonetics*, *31*(3–4), 305–320. https://doi.org/10.1016/S0095-4470(03)00030-5

Gordon, M. K. (2016). *Phonological Typology*. Oxford: Oxford University Press.

Goslin, J., & Frauenfelder, U. H. (2001). A comparison of theoretical and human syllabification. *Language and Speech*, *44*(4), 409–436. https://doi.org/10.1177/00238309010440040101

Hall, T. A. (2006). English syllabification as the interaction of markedness constraints. *Studia Linguistica*, *60*(1), 1–33. https://doi.org/10.1111/j.1467-9582.2006.00131.x

Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics*, *31*(3–4), 373–405. https://doi.org/10.1016/j.wocn.2003.09.006

Hay, J., Pierrehumbert, J., & Beckman, M. (2004). Speech perception, well-formedness and the statistics of the lexicon. In J. Local, R. Ogden, & R. Temple (Eds.), *Papers in Laboratory Phonology VI* (pp. 58–74). Cambridge: Cambridge University Press.

Kahn, D. (1976). *Syllable-based generalizations in English phonology [dizertační práce]*. Cambridge, Mass.: MIT.

Kingston, J., Kawahara, S., Chambless, D., Mash, D., & Brenner-Alsop, E. (2009). Contextual effects on the perception of duration. *Journal of Phonetics*, *37*(3), 297–320. https://doi.org/10.1016/j.wocn.2009.03.007

Machač, P., & Skarnitzl, R. (2009). *Principles of Phonetic Segmentation*. Praha: Epocha.

Maddieson, I. (1985). Phonetic cues to syllabification. In V. Fromkin (Ed.), *Phonetic Linguistics: Essays in Honor of Peter Ladefoged* (pp. 203–221). New York: Academic Press.

Munson, B. (2001). Phonological pattern frequency and speech production in adults and children. *Journal of Speech, Language, and Hearing Research*, *44*(4), 778–792.

Ní Chiosáin, M., Welby, P., & Espesser, R. (2012). Is the syllabification of Irish a typological exception? An experimental study. *Speech Communication*, *54*(1), 68–91. https://doi.org/10.1016/j.specom.2011.07.002

Parker, S. (2008). Sound level protrusions as physical correlates of sonority. *Journal of Phonetics*, *36*(1), 55–90. https://doi.org/10.1016/j.wocn.2007.09.003

Port, R. (2007). How are words stored in memory? Beyond phones and phonemes. *New Ideas in Psychology*, *25*(2), 143–170. https://doi.org/10.1016/j.newideapsych.2007.02.001

Pulgram, E. (1970). *Syllable, Word, Nexus, Cursus*. The Hague: Mouton.

R Core Team. (2016). R: A language and environment for statistical computing (Version 3.2.4). Vienna: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Redford, M. A., & Randall, P. (2005). The role of juncture cues and phonological knowledge in English syllabification judgments. *Journal of Phonetics*, *33*(1), 27–46. https://doi.org/10.1016/j.wocn.2004.05.003

Schiller, N. O., Meyer, A. S., & Levelt, W. J. M. (1997). The syllabic structure of spoken words: evidence from the syllabification of intervocalic consonants. *Language and Speech*, *40*, 103–140.

Sendlmeier, W. F. (1995). Feature, phoneme, syllable or word: how is speech mentally represented? *Phonetica*, *52*(3), 131–143.

Šturm, P. (2017). *Určování slabičných hranic v češtině [dizertační práce]*. Praha: Filozofická fakulta Univerzity Karlovy.

Šturm, P., & Lukeš, D. (2017). Fonotaktická analýza obsahu slabik na okrajích českých slov v mluvené a psané řeči. *Slovo a slovesnost*, *78*(2), 99–118.

Treiman, R., & Danis, C. (1988). Syllabification of intervocalic consonants. *Journal of Memory and Language*, *27*(1), 87–104. https://doi.org/10.1016/0749-596X(88)90050-2

Treiman, R., Kessler, B., Knewasser, S., Tincoff, R., & Bowman, M. (2000). English speakers' sensitivity to phonotactic patterns. In N. B. Broe & J. Pierrehumbert (Eds.), *Laboratory Phonology V: Acquisition and the Lexicon* (pp. 269–282). Cambridge: Cambridge University Press.

Vitevitch, M. S., Luce, P. A., Charles-Luce, J., & Kemmerer, D. (1997). Phonotactics and syllable stress: Implications for the processing of spoken nonsense words. *Language and Speech*, (40), 47–62.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer.

Zec, D. (2007). The syllable. In P. de Lacy (Ed.), *The Cambridge Handbook of Phonology* (pp. 161–194). Cambridge: Cambridge University Press.

APPENDIX

Temporal structure of the target stimuli (V = first vowel, C1 = first consonant in the intervocalic cluster, C2 = second consonant). Onset frequency was adopted from Šturm and Lukeš (2017), and refers to the ipm (items per million) frequency of occurrence of the cluster as a word-initial onset in written texts.

sonority	cluster	word	onset freq.	V dur (ms)	C1 dur (ms)	C2 dur (ms)	C1/C2 ratio					C1/V ratio	C2/V ratio
							unchanged	C1 long	C2 short	C1 short	C2 long	unchanged	unchanged
O-O	/sk/	viskem	2 100	67	163	95	1.72	2.57	3.43	0.86	1.14	2.43	1.42
O-O	/st/	lesta	8 371	75	192	99	1.94	2.91	3.88	0.97	1.29	2.56	1.32
O-O	/xt/	zachty	0	82	122	135	0.90	1.36	1.81	0.45	0.60	1.49	1.65
O-O	/kt/	čaktem	6 380	71	100	133	0.75	1.13	1.50	0.38	0.50	1.41	1.87
O-O	/tk/	natka	15	82	133	105	1.27	1.90	2.53	0.63	0.84	1.62	1.28
O-O	/t͡sk/	smacky	0	70	205	115	1.78	2.67	3.57	0.89	1.19	2.93	1.64
O-S	/br/	chebra	1 049	67	111	84	1.32	1.98	2.64	0.66	0.88	1.66	1.25
O-S	/tr/	knetrem	1 748	80	123	51	2.41	3.62	4.82	1.21	1.61	1.54	0.64
O-S	/sl/	keslo	2 950	88	166	90	1.84	2.77	3.69	0.92	1.23	1.89	1.02
S-S	/rm/	charmu	0	102	98	110	0.89	1.34	1.78	0.45	0.59	0.96	1.08

RESUMÉ

Cílem studie bylo ověřit, zda akustický signál obsahuje percepčně relevantní vodítka ohledně dělení slov na slabiky. Sylabifikační odezvy 27 českých mluvčích byly zkoumány v behaviorálním experimentu za použití dvojslabičných pseudoslov s deseti shluky CC. Poměr trvání C1 a C2 intervokalického shluku byl manipulován čtyřmi způsoby: zkrácením/prodloužením prvního/druhého konsonantu. Úkolem mluvčích bylo opakovat slyšené stimuly po slabikách (metoda vkládání pauz). Logistická regrese potvrdila signifikantní efekt sonoritního typu shluku, fonotaktiky na okrajích slov a sylabifikační strategie (jen polovina účastníků uvedla, že se v experimentu neřídila žádnou strategií). Typ manipulace se na jednu stranu neprojevil jako signifikantní prediktor v logistické analýze; na druhou stranu poměr C1/C2 v souladu s hypotézou negativně koreloval s mírou rozdělení shluku. Tato korelace naznačuje, že čím je C1 delší než C2, tím vyšší má shluk pravděpodobnost, že bude zachován jako prétura následující slabiky.

Pavel Šturm
Institute of Phonetics
Faculty of Arts, Charles University
pavel.sturm@ff.cuni.cz

ACTA UNIVERSITATIS CAROLINAE
PHILOLOGICA 3/2017

Editors: doc. Mgr. Radek Skarnitzl, Ph.D.
doc. PhDr. Jan Volín, Ph.D.
Cover and layout by Kateřina Řezáčová
Published by Charles University
Karolinum Press, Ovocný trh 560/5, 116 36 Praha 1
www.karolinum.cz
Prague 2017
Typeset by Karolinum Press
Printed by Karolinum Press
ISSN 1804-624X (Print)
ISSN 2464-6504 (Online)
MK ČR E 19831

Distributed by Faculty of Arts, Charles University,
2 Jan Palach Sq., 116 36 Prague 1, Czech Republic
(books@ff.cuni.cz)