ACTA UNIVERSITATIS CAROLINAE
PHILOLOGICA 1/2022

ACTA UNIVERSITATIS CAROLINAE

PHILOLOGICA 1/2022

JAN VOLÍN and PAVEL ŠTURM (eds.)

CHARLES UNIVERSITY
KAROLINUM PRESS
2022

Editors: Jan Volín (Charles University)
Pavel Šturm (Charles University)

http://karolinum.cz/journals/philologica

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

CONTENTS

EDITORIAL

The first issue of Acta Universitatis Carolinae dedicated to phonetics – *Phonetica Pragensia* – was published 55 years ago, in 1967. It was the year when the International Congress of Phonetic Sciences (ICPhS) was organized in Prague because Czech phoneticians were recognized by the International Phonetic Association as significant contributors to the development of the scientific field. The phonetic issue of AUC of that year was dedicated to this world event.

The current issue honours the legacy of the previous generations of Czech phoneticians. After more than a hundred years of its existence, the Prague Institute of Phonetics stands on firm grounds. Its team consists of seven active researchers and educators, who, for the sake of historical comparison, can be listed as follows: one Professor Emeritus, one Full Professor, one Associated Professor, three Assistant Professors with a doctoral degree, and one part-time Assistant with an MA degree. The members of the team maintain many domestic and international friendships and some of the external colleagues contributed to this issue as authors, others performed as reviewers, and many more awaited this issue keenly to read it.

The work of the current team is widely recognized and appreciated, too. This can be documented by the fact that the Institute of Phonetics in Prague was bestowed the honour to organize the 4th International Workshop on the History of Speech Communication Research in 2021 (an ISCA event), the annual conference of the International Association for Forensic Phonetics and Acoustics (IAFPA) in 2022, and the 20th International Congress of Phonetic Sciences in 2023.

The forensic area is represented in the current issue by the first two articles. Tomáš Nechanský and his colleagues investigated the perceptual impact of mismatched recordings. Their large sample (300 recordings from 100 speakers) represented both language and time mismatch, i.e., an identical speaker was recorded speaking two different languages and, also, was repeatedly recorded at different times. The perceptual consequences for human listeners are contrasted in this study with the artificial intelligence achievement in speaker identification. A similar problem is tackled in the contribution by Maral Asiaee and Homa Asadi from Iran. They worked with bilingual users of Persian and Sorani Kurdish to see whether certain salient acoustic features pertinent to speaker's voice characteristics change when an individual switches from one language to another.

The third study in this issue investigated the perceptual effects of the presence or absence of glottal stops before word-initial vowels in Czech conversational speech. The authors, Michaela Svoboda and Pavel Šturm, used authentic political TV debates as the source of their material. In the design of correlated samples, they either removed or added glottal stops into the utterances from the debates and asked listeners to react to the utterances in a reaction-time perception test. The results point at facilitative function of the glottal stop, but also at complex interactions within the linguistic contents.

https://doi.org/10.14712/24646830.2021.10

The article submitted by Lauri Tavi from the University of Eastern Finland is also concerned with the perception of speech, even if it is investigated by proxy. The author is interested in the impact of fast speech on prosodic forms of utterances. Specifically, he is measuring acoustic features reflecting the decrease of accent prominence when a person is compelled to speak faster. The proposed metric – *syllabic prosody index* (SPI) – is definitely worth noting. It could be expected to have further utilization in speech prosody research. Needless to say, fast speech is confirmed to be less intelligible to an automatic speech recognition system.

Speech tempo is the subject matter of the fifth contribution, too. Rather than asking respondents to speak slow and then fast, Jan Volín works with a set of recordings in which some of the speakers are habitually slow or fast. Two speech genres are investigated: news reading and poetry reciting. The author examined four types of variation in articulation and speech rates, and he correlated the measures to prosodic phrasing. One of the research questions was whether the speakers who produce habitually faster speech also make fewer prosodic boundaries than the slower speakers do.

Although the previous two studies focus on speech tempo, they contribute quite significantly to the clarification of methodological issues in phonetics. Indeed, the matter of research methodology is a perpetual concern, and two studies are directly dedicated to it. Alžběta Houzar and Radek Skarnitzl test three methods of extraction of vowel formants in two types of material: spontaneous speech and read-out sentences. They make an effort to determine how the three methods capture inter-speaker and intra-speaker variability. The results indicate quite convincingly that researchers need to be cautious when interpreting formant values obtained by different methods.

Speech melody is a salient phenomenon that has occupied the minds of linguists for a very long time. Yet, even in this area, research methodology is still debatable and a search for more adequate methods is imperative. One of the major concerns here is the clear link between quantitative precision and common linguistic concepts. Michaela Svatošová and Jan Volín are offering their contribution to the problem in the seventh article of the present AUC issue. They explain and advocate the use of Legendre polynomials for the description of traditionally recognized Czech melodemes (nuclear patterns). Their proposal, too, could be of interest to researchers internationally.

Last but not least, the current AUC issue brings two papers concerning speech acquisition. Šárka Šimáčková and Václav Jonáš Podlipský examined the effect of an online general pronunciation course on prosodic skills of adult EFL learners. Those skills were represented by pitch span (standing for liveliness of speech performance) and tempo (standing for professionalism in delivery). The outcome of the experiment is quite encouraging. It suggests that even online courses can have a measurable effect on students.

Learners of foreign languages not only acquire new speech production skills, but they can also lose certain pronunciation patterns of their mother tongue in a process called attrition. This is a topic of the study by Marie Hévrová and Tomáš Bořil. They worked with recordings of late Czech-French bilinguals and performed extensive analyses of various vowels and two fricative consonants. Indeed, late exposure to French influenced the speech characteristics of Czech, the mother tongue of the subjects who participated in the study.

As apparent from the previous paragraphs, the individual contributions in this issue of AUC journal are not alphabetically ordered. We arranged them by topic proximity with the aim to create a thematic flow which could facilitate the reader's appreciation of mutual links between various challenges in phonetic research. It is our sincere desire that the readers enjoy the variety of topics and the quality of research in this issue of AUC journal.

Jan Volín and Pavel Šturm
https://doi.org/10.14712/24646830.2022.24

2022 ACTA UNIVERSITATIS CAROLINAE PAG. 11–22
PHILOLOGICA 1 / PHONETICA PRAGENSIA

THE IMPACT OF MISMATCHED RECORDINGS ON AN AUTOMATIC-SPEAKER-RECOGNITION SYSTEM AND HUMAN LISTENERS

TOMÁŠ NECHANSKÝ, TOMÁŠ BOŘIL, ALŽBĚTA HOUZAR, RADEK SKARNITZL
Institute of Phonetics, Faculty of Arts, Charles University

ABSTRACT

The so-called 'mismatch' is a factor which experts in the forensic voice comparison field encounter regularly. Therefore, we decided to explore to what extent the automatic-speaker-recognition system's and the earwitness' ability to identify speakers is influenced when recordings are acquired in different languages and at different times. 100 voices in a database of 300 recordings (100 speakers recorded in three mutually mismatched sessions) were compared with an automatic-speaker-recognition software VOCALISE based on i-vectors and x-vectors, and by 39 respondents in simulated voice parades. Both the automatic and perceptual approach seem to have yielded similar results in that the less complex the mismatch type, the more successful the identification. The results point to the superiority of the x-vector approach, and also to varying identification abilities of listeners.

Keywords: forensic voice comparison, temporal mismatch, language mismatch, automatic speaker recognition, voice parade

1. Introduction

Forensic phoneticians, when identifying a speaker, encounter various cases differing in complexity. Even two realizations of the same word uttered right after each other will not be identical from the acoustic point of view, and this variability in speech must be acknowledged when comparing voices for forensic purposes. A ubiquitous characteristic of Forensic Voice Comparison (FVC) which increases the complexity of the process is *mismatch* between recordings, which can take several forms.

First, recordings examined in FVC typically differ in their technical aspects, particularly in the characteristics of the channel. For example, voice samples of the unknown speaker (typically the perpetrator) may originate from an intercepted mobile telephone call or from a wiretapped office, while those of the known speaker (the suspect) may be obtained in an interrogation room. The effect of channel variation has been investigated by numerous researchers, with special focus on telephone transmission (both landline and

https://doi.org/10.14712/24646830.2022.25

mobile); results of such studies are not surprising, with mismatched recordings yielding lower recognition scores (e.g., Alexander et al., 2005; Hughes et al., 2019; Bortlík, 2021). Technical mismatch also includes phenomena such as reverberation or various forms of background noise (see Guillemin, 2022 for a comprehensive summary).

Speakers themselves constitute sources of several kinds of mismatch. This kind of within-speaker variation stems from the incredible plasticity of our speech production mechanism. Some of the areas which have received considerable attention include various speech styles and the effect that they have on specific acoustic parameters (Jessen, 2009; McDougall & Duckworth, 2018; Ross et al., 2019), the impact of various affective or physiological states (Eriksson et al., 2007; Scherer, 2019), as well as phonetic accommodation to one's communication partner (see, e.g., Earnshaw, 2021; Šturm et al., 2021). In FVC, these are particularly important since the analyzed recordings tend to be mismatched in this respect. In general, research points to differences in the values of acoustic parameters and sometimes to decreased speaker recognition performance (e.g., Shriberg & Scheffer, 2009). However, some studies suggest that certain parameters remain relatively stable within speakers. For example, McDougall and Duckworth (2018) report considerable within-speaker consistency in various dysfluency features in telephone and interview styles. Other behavioural effects, discussed in more detail by Gold et al. (2022), include whispered speech, loud speech in the presence of Lombard effect, as well as disguised speech (Eriksson, 2010; Růžičková & Skarnitzl, 2017).

Another source of mismatch consists in the non-contemporary nature of FVC: the recordings which are compared in a forensic case must have been, by definition, obtained at different times. This has been examined by a number of researchers from the forensic-phonetic, as well as automatic speaker recognition (ASR) perspective. Their studies focused on the recognizability of speakers across different time spans, from several days (Ross et al., 2019; what is often referred to as 'session mismatch') and months (Kelly & Hansen, 2015) to years or even several decades (Hollien & Schwartz, 2000; Rhodes, 2017). It is not surprising that speech and voice patterns change throughout our lifetime, resulting in a drop of both human and machine recognition of speakers. For example, Rhodes' (2017) investigation of speakers over a span of 28 years showed a change in vowel formants of between 3 and 15%, with the most robust effect observed in *F*1. Likelihood ratios obtained from vowel formant data shifted towards incorrect decisions, and ASR performance dropped significantly at delays between recordings above 14 years.

The final source of within-speaker variability to be mentioned here is when different languages or accents are used by one speaker. Several studies have addressed foreign accent in FVC, focusing on the imitation of a foreign accent (Torstensson et al., 2004), on listeners' ability to identify authentic foreign accents (Neuhauser & Simpson, 2007; Sullivan & Schlichting, 2000), or on the degree to which non-native language background helps witness experts identify a speaker of another language (Schiller et al., 1997). Language-mismatched recordings have also been examined using ASR approaches. For example, Misra and Hansen (2014) found that when only English recordings were used for training the ASR system, language mismatch resulted in a drop in performance by a factor of 2.5; however, including non-English material at the training stage substantially improved performance. In a recent study, Bortlík (2021) examined the effect of foreign-accented speech on the performance of state-of-the-art ASR systems; he reported higher error rates

in language-mismatched comparisons – i.e., when a Czech speaker was speaking Czech in one condition and English in the other – than in matched comparisons.

The first aim of the present study is to investigate the combined effect on the performance of an ASR system of contemporary and non-contemporary recordings of speech produced in the same and in a foreign language, by the same speaker. The setting simulates two hypothetical situations which may be relevant for FVC. During the perpetration of a crime, the unknown speaker uses a foreign language (L2), while the suspect recording with the known speaker is in the speaker's mother tongue (L1). Since it is not unusual for the suspect recording to originate from a wiretap, language identity cannot be ensured, and cross-language comparisons will be required. The second aim is to explore the ability of listeners to identify the speaker in a simulated voice parade under the conditions described above.

2. Method

2.1 Material

The database for our research comprises 100 (78 female and 22 male, aged 20–25) speakers of Czech as L1 and English as L2 (with the CEFR level being B2 or C1). The speakers were studying English and American Studies at Charles University at the time of recording. The recordings were obtained in the sound-treated recording studio of the Institute of Phonetics in Prague, using the high-quality AKG C4500 B-BC condenser microphone, with 32-kHz sampling frequency and 16-bit resolution.

Three recording sessions are analyzed in this study. At the beginning of their studies, the speakers were asked to read: first, a phonetically rich text in Czech; and second, several pieces of BBC news in English. Four months later, the same students were recorded reading other BBC news texts in English again. On average, each participant produced ca. 1 minute of speech in Czech, and 3–4 minutes in English twice.

2.2 Automatic speaker recognition procedure

Since all the speakers are known but were recorded under three conditions, for each speaker we compared the following mismatches as if they were the unknown and suspect recordings:

- language mismatch (contemporary Czech and English recordings),
- temporal mismatch (non-contemporary English recordings),
- double mismatch (non-contemporary Czech and English recordings).

Speaker comparisons were performed in VOCALISE by Oxford Wave Research, using the i-vector (session VOCALISE i-vector 2017B) and x-vector (session VOCALISE x-vector 2019A-Beta-RC2) PLDA framework. In this framework, vectors of speakers (i-vector or x-vector) are compared using probabilistic linear discriminant analysis (PLDA); this post-processing method computes the likelihood of the vector pair originating from the same speaker versus coming from two different speakers. The i-vectors and x-vectors are different ways of speaker modelling in the speaker recognition pipeline. Whereas i-vectors make use of front-end factor analysis as the feature extractor, x-vectors rely on trained deep neural

networks (see Kelly et al., 2019 for more details); x-vectors are the most recent approach to speaker modelling. The resulting scores were calibrated using cross-validation in the BioMetrics software by Oxford Wave Research.

Apart from the three comparisons listed above, we conducted several partial comparisons to examine the effect of "tuning" (see Skarnitzl et al., 2019) using condition adaptation. Condition adaptation optimizes the ASR system to new conditions, specific to the given recordings, by adapting the LDA and PLDA models. By performing condition adaptation, the properties from dozens of i-/x-vectors in the adaptation set are used to adapt tens of thousands of i-/x-vectors in the training dataset of VOCALISE towards the new conditions; in other words, the statistics of the LDA and PLDA models were updated using a weighted combination of the original training data and the recordings provided by the authors. For this purpose, the three datasets were divided into two halves (50 speakers in set 1 and 50 in set 2; the division was random but identical across the three datasets). Subsequently, recordings of set-2 speakers were used for condition adaptation of set-1 comparisons, and vice versa.

We will report the equal error rate (EER) as the standard measure of ASR performance (EER is defined as the number when false-acceptance rate and false-rejection rate become equal; see Hansen & Hasan, 2015). Since some comparisons involve relatively small datasets, Convex Hull EER values are reported in all analyses. In addition, we will provide values of the log-likelihood-ratio cost (C_{llr}), a measure that evaluates the accuracy of an ASR system by capturing the gradient goodness of a set of likelihood ratios derived from test data, with values ideally not exceeding 1 (Morrison, 2011).

2.3 Listening test procedure

For our perception experiment – a simulated voice line-up in which an earwitness is supposed to identify the perpetrator's voice among recordings of suspects, we used recordings of 22 male speakers from the same database. Each line-up (or parade) featured six recordings: the perpetrator's voice and five suspects' voices available for matching with the perpetrator. Crucially, to approximate conditions of real-life voice parades, the perpetrator's voice could be either present among the five suspects (i.e., the so-called target voice), or absent. The perpetrator and suspect recordings (whether the target was present or absent) would differ in language (language mismatch), time of recording (temporal mismatch), both (double mismatch), or would not differ at all (no mismatch). The perception experiment comprised the following line-up conditions:

- 5 line-ups for recordings of no mismatch (contemporary Czech, or English),
- 4 line-ups for recordings of language mismatch (contemporary Czech and English),
- 2 line-ups for recordings of temporal mismatch (non-contemporary English),
- 4 line-ups for recordings of double mismatch (non-contemporary Czech and English).

In real-life voice parades, it is recommended that foils' voices (i.e., all suspect voices except the target) should be as similar to the target speaker's voice as possible (de Jong-Lendle et al., 2015). We used fundamental frequency (f_0) median as a measure of distance (i.e., similarity) between speakers. We selected the foils' voices to be closest to the perpetrator. This was not adhered to only when the target's and perpetrator's samples were mismatched; in this case, the most distant speakers were chosen.

In total, 90 samples (15 line-ups * 6 samples) of about 5 seconds in duration were used for the perception test. It was ensured that the samples within one line-up were not textually identical and that they were loudness-normalized. The perception test was designed in PsyToolkit (Stoet, 2010; 2017) and was coded as fifteen tasks requiring the participant to first listen to the perpetrator, then to the five suspects, and after that to either match one of the suspects with the perpetrator, or to check a box indicating that the perpetrator's voice was not one of the suspects. Participants were allowed to replay any of the recordings. In order to minimize the order effect, samples in each line-up, as well as the line-ups themselves were randomized.

Besides the experiment, we gathered basic demographic information from the respondents, who received monetary compensation for their participation. The perception experiment was completed by 33 female and 7 male respondents, aged 22–49, all coming from the Czech Republic. It was revealed later that one participant had not listened to the stimuli properly, and her responses were eliminated. Therefore, 39 listeners' answers were analyzed in the end. The total time spent ranged from 15 to 46 minutes; 80% of the participants finished the test (including the demographic survey) under 31 minutes.

3. Results

3.1 Automatic speaker recognition

Figure 1 compares equal error rates achieved by the i-vector and x-vector approaches: it is obvious (notice the difference in the scale of the two plots) that the i-vectors are significantly outperformed by the x-vectors.

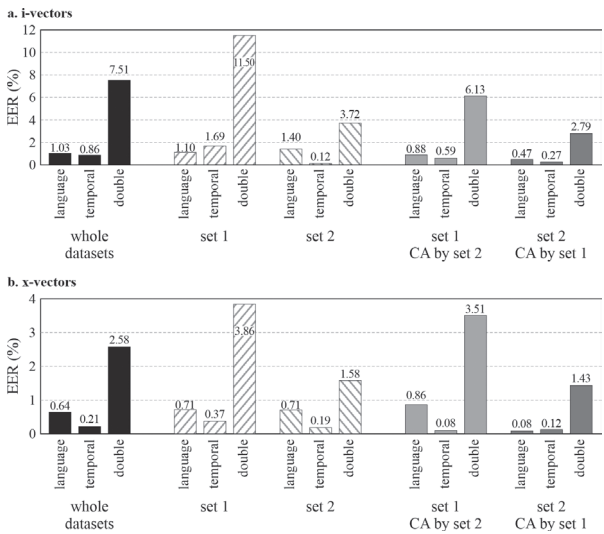

Figure 1. Equal error rates for i-vectors (a.) and x-vectors (b.) for three types of mismatch (language, temporal, double). In the left part (black), results for the entire datasets; in the middle (with stripes), for the half-size datasets; on the right (in grey), for the half-size datasets with the opposing half used for condition adaptation (CA).

What is particularly noteworthy is that single mismatch conditions (i.e., only temporal, or only language mismatch) result in very good performance, with EERs around 1% or lower, using both i-vectors and x-vectors. However, double mismatch conditions (both non-contemporary and language-mismatched) yield significantly higher error rates, 7.5% for i-vectors and 2.6% for x-vectors. The situation may also be illustrated using a combined equal error plot, with all three main comparisons (see Fig. 2); note that only results for x-vectors are shown in the figure. An equal error plot shows the false acceptance rate (FAR) and the false rejection rate (FRR) on the vertical axis against the threshold score on the horizontal one; the intersection of the two curves corresponds to the EER. The better the curves are separated, the better the recognition.

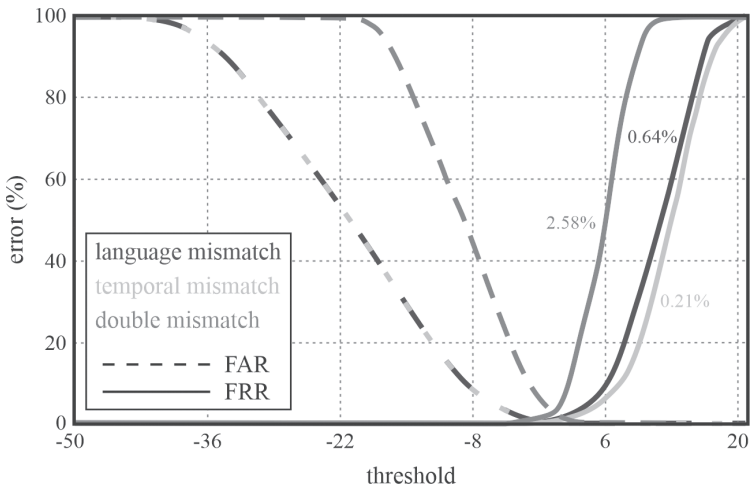

Figure 2. Equal error plot comparing the three main comparisons using x-vectors; EER values are shown in the corresponding colours. FAR = false acceptance rate, FRR = false rejection rate

The same tendencies can be observed also for the comparisons of partial datasets (shown with stripes in Fig. 1) and for partial datasets "tuned" by the corresponding opposing half using condition adaptation (in shades of grey). At the same time, EER clearly depends on the particular selection of speakers under comparison: results for set 1 and set 2 are far from identical. System accuracy can be regarded as high for all comparisons performed, with C_{llr} being 0.4 for the double-mismatched condition in i-vector set 1, and considerably lower in all other (i-vector and x-vector) comparisons ($C_{llr} < 0.1$ for all single-mismatched comparisons, and $0.07 \leq C_{llr} \leq 0.4$ for the double-mismatched conditions).

What remains to be discussed is the hypothesized benefit of condition adaptation on ASR performance; in other words, we are interested in finding out whether using the opposing half of the dataset for PLDA adaptation yields lower error rates. Overall, this benefit was slightly more salient in the case of i-vectors, where we can see a considerable improvement in most of the scores (cf. the striped and grey bars in Fig. 1a); in the case of x-vectors, condition adaptation yielded a lower EER in five out of the six comparisons (Fig. 1b). Crucially, the benefit turned out to be greatest with the double-mismatched conditions.

3.2 Listening test

First, we wanted to know how listeners scored individually and what the overall successful identification rate was. The results for individual participants are presented in Figure 3 for the target-present scenario and in Figure 4 for the target-absent scenario. The figures provide overviews of hits (correctly identified targets), foils (incorrectly identified suspects), correct rejections (correctly rejected all suspects), and incorrect rejections (incorrectly rejected all suspects).

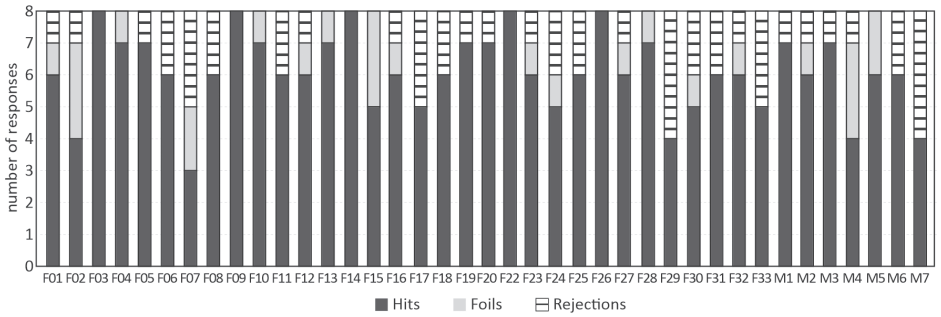

Figure 3. Individual responses (hits, foils, and incorrect rejections) for target-present parades (see text).

Since there were 8 target-present line-ups (see Fig. 3), the maximum number of correct answers (labelled as hits in the figure) was 8, which was achieved by five listeners. On the other hand, this condition allowed for two types of mistakes – incorrectly identifying another suspect as the target (labelled as foils) and incorrectly rejecting all suspects (assuming the target was absent; marked as rejections). The highest number of incorrect answers combined was 5, which was produced by only one listener (i.e., a successful identification rate of only 37.5%).

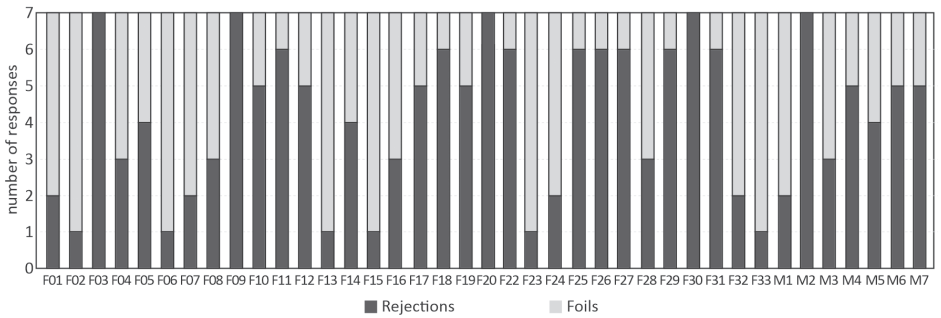

Figure 4. Individual responses (correct rejections and foils) for target-absent parades (see text).

As for target-absent parades (Fig. 4), it was possible to either answer correctly (rejecting all suspects, as the target was not included; labelled as rejections in the figure), or to incorrectly choose a suspect (foils). We had 7 of such parades, and a 100% successful identification rate was again achieved by five listeners (of which two also scored 100% in

the target-present setup). Overall, for both scenarios, out of 15 parades 61.5% of listeners managed to solve 10 or more, 38.5% solved 12 or more correctly, and only 2 listeners (0.5%) succeeded in all.

Second, we were interested in whether any of the collected demographic data corresponded with identification rates; however, none of respondents' sex, age, education, nor level of English proved significant. Note that education was treated as a binary variable, with participants either with or without linguistic background.

Third, we explored whether respondents were able to perform better (i.e., scored a higher number of correct answers) in a specific type of mismatch. To find out, all parades were divided into target-present and target-absent groups and according to mismatch type (none, temporal, language, double). The results are shown in Table 1.

Table 1. Percentages of all correct answers for all line-up types. * marks comparisons with the no-mismatch target-present condition (°) which turned out significant ($p < 0.05$).

TARGET-PRESENT SCENARIO	
mismatch type	**correct answers**
none °	98.3%
temporal	79.5%
language *	60.3%
double *	56.4%

TARGET-ABSENT SCENARIO	
mismatch type	**correct answers**
none	66.7%
temporal	64.1%
language *	61.5%
double *	46.2%

To assess the statistical significance of the reported relationships, we used R (R Core Team, 2022) and the *lme4* (Bates et al., 2015) and *afex* (Singmann et al., 2022) packages to perform a mixed effects logistic regression analysis (bobyqa optimizer) of correct and incorrect responses (correct responses include hits and correct rejections). As fixed effects, we entered Target and Mismatch with an interaction term into the model. As random effects, we used intercepts for Suspect and Participant, as well as by-Participant random slopes for the effect of Target. P-values were obtained by performing pairwise post-hoc tests with Tukey method of p-value adjustment for comparing a family of 8 estimates using the *emmeans* package (Lenth, 2022). We found significant differences ($p < 0.05$) between target-present line-ups without any mismatch (marked ° in Table 1) and the four parade groups which are marked with an asterisk in the table.

Finally, we wanted to see whether it is true that the more times the listener played recordings in a parade, the more likely it was for them to answer correctly. As Figure 5 reveals, this is not the case. For target-absent line-ups, no correlation was found (Pearson correlation coefficient $r = 0.18$; $p = 0.707$). On the other hand, for target-present parades, we discovered a strong negative correlation between the number of playbacks and correct answers ($r = -0.92$; $p = 0.00138$). In other words, repeated playback of the voices in the parade did not result in higher accuracy of the listeners; on the other hand, it strongly correlated with the listeners' decision-making uncertainty in line-ups featuring a lower successful identification rate.

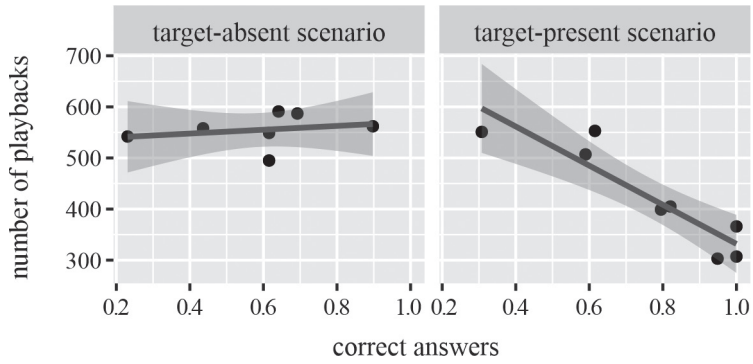

Figure 5. Relationship between the number of replayed recordings and correct answers.

4. Discussion

In this paper, we aimed to find whether mismatched recordings have any impact on speaker identification performance of an ASR system and of human listeners. Our data clearly prove, in line with previous research, that once voices come from mismatched sources, performance does worsen.

As for ASR, we expected that the global validity of the system would decline with increasing "dissimilarity" of the compared datasets. Since the span between the English recordings was only four months, we believed that the speakers' already advanced level of English had not improved considerably, and thus their production remained similar; nevertheless, it was certainly possible for the temporally mismatched recording to have been affected by other changes, such as illness. On the other hand, we supposed that speakers had used different phonetic settings and produced acoustically different phones and prosody in Czech and English. Then, it seemed logical that the combination of these two mismatches would reflect in the results.

For the original datasets, divided datasets, and adapted datasets comparisons (by both i- and x-vectors), our assumptions were confirmed although the difference in EER between language and temporal mismatch was relatively small. The exceptions are set 1 under i-vectors and set 2 calibrated with set 1 under x-vectors when VOCALISE performed slightly better in language than temporal mismatch. As discussed in Section 3.1, system tuning by means of condition adaptation did not always turn out to be beneficial.

Regarding our perception experiment, we added a no-mismatch condition (representing the most similar type on our scale) and wondered whether listeners as well would confirm our "dissimilarity" hypothesis stated above. Even though the percentages of correct answers seemed promising (see Tab. 1), statistically we were able to confirm only a fraction of significant pairs. It would be interesting to replicate the experiment with a higher number of respondents to establish that the more complex the mismatch is, the less successful in speaker identification people are.

It is worth mentioning that there were considerable identification differences amongst individual participants. Whereas we witnessed two "super-recognizers" who solved all 15 parades, there were seven respondents who responded incorrectly in more than 50%

of line-ups. It is worth pointing out that the two successful participants are university students of phonetics; and one of them managed to complete the experiment in 15 minutes (compared to the mean of 26.3) without even having to listen to all suspect recordings in six of the fifteen line-ups.

There were more listeners who did not need to listen to all suspect voices: over 56% of respondents correctly chose a target in at least one parade in this manner (and 28% in two or more parades). Clearly, the experiment featured speakers whose voice characteristics were so salient that it was not necessary for the respondent to continue listening to others. In fact, two speakers were identified in 100% and one in 95% of cases. Conversely, there were two speakers that were much more difficult to recognize – only in 23% and 31% of cases. Also, we registered a suspect speaker who was incorrectly identified as the perpetrator in 37 cases in the target-absent scenario. Instrumental analyses of these speakers' voices could reveal further details as to why he is easily mistaken for other speakers; however, this is beyond the scope of this study.

To conclude, we have shown that ASR systems perform noticeably worse when analyzing voices recorded in different languages and at different times. Nevertheless, in our perception experiment the listeners' ability to identify the perpetrator also dropped considerably as compared to recognizing matched voices. The comparison of known and unknown recordings originating from mismatched sources is far from trivial and it is something of which forensic experts, when drawing conclusions, should be aware.

Acknowledgements

This study was supported by the project the Grant Schemes at CU, reg. no. CZ.02.2.69/0.0/0.0/19_073/0016935.

REFERENCES

Alexander, A., Dessimoz, D., Botti, F., & Drygajlo, A. (2005). Aural and automatic forensic speaker recognition in mismatched conditions. *International Journal of Speech, Language and the Law*, 12(2), 214–234.

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

Bortlík, J. F. (2021). *Czech accent in English: Linguistics and biometric speech technologies*. Palacký University Olomouc. (unpublished PhD dissertation)

de Jong-Lendle, G., Nolan, F., McDougall, K., & Hudson, T. (2015). Voice lineups: A practical guide. In: *Proceedings of ICPhS 2015*, paper 0598.

Earnshaw, K. (2021). Examining the implications of speech accommodation for forensic speaker comparison casework: A case study of the West Yorkshire Face vowel. *Journal of Phonetics, 87*, 101062.

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 University Press.

Eriksson, E. J., Rodman, R. D., Hubal, R. C. (2007). Emotions in speech: Juristic implications. In: Müller, C. (Ed.), *Speaker classification I* (pp. 152–173). Springer-Verlag.

Gold, E., Ross, R., & Earnshaw, K. (2022, accepted). Within-speaker variation: Speaker-based causes. In: Nolan, F., McDougall K., & Hudson, T. (Eds.), *Oxford handbook of forensic phonetics*. Oxford University Press.

Guillemin, B. (2022, accepted). Within-speaker variation: External causes. In: Nolan, F., McDougall K., & Hudson, T. (Eds.), *Oxford handbook of forensic phonetics*. Oxford University Press.

Hansen, J. H. L., & Hasan, T. (2015). Speaker recognition by machines and humans: A tutorial review. *IEEE Signal Processing Magazine, 32*(November), 74–99.
Hollien, H., & Schwartz, R. (2000). Aural-perceptual speaker identification: Problems with noncontemporary samples. *Forensic Linguistics*, 7(2), 199–211.
Hughes, V., Harrison, P., Foulkes, P., French, P., & Gully, A. J. (2019). Effects of formant analysis settings and channel mismatch on semi-automatic forensic voice comparison. In: *Proceedings of ICPhS 2019*, 3080–3084.
Jessen, M. (2009). Forensic phonetics and the influence of speaking style on global measures of fundamental frequency. In: Grewendorf, G., & Rathert, M. (Eds.), *Formal linguistics and law* (pp. 115–139). Mouton de Gruyter.
Kelly, F., Forth, O., Kent, S., Gerlach, L., & Alexander, A. (2019). Deep neural network based forensic automatic speaker recognition in VOCALISE using x-vectors. Presented at the *Audio Engineering Society (AES) Forensics Conference 2019*, Porto, Portugal, 2019. Retrieved from https://www.aes.org/e-lib/browse.cfm?elib=20477
Kelly, F., & Hansen, J. H. L. (2015). Evaluation and calibration of short-term aging effects in speaker verification. In: *Proceedings of Interspeech 2015*, 224–228.
Lenth, R. (2022). emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.7.4-1, <https://cran.r-project.org/package=emmeans>.
McDougall, K., & Duckworth, M. (2018). Individual patterns of disfluency across speaking styles: A forensic phonetic investigation of Standard Southern British English. *International Journal of Speech, Language and the Law*, 25(2), 205–230.
Misra, A., & Hansen, J. H. L. (2014). Spoken language mismatch in speaker verification: An investigation with NIST-SRE and CRSS BI-LING corpora. In: *Proceedings of the IEEE Spoken Language Technology Workshop*, 372–377.
Morrison, G. S. (2011). Measuring the validity and reliability of forensic likelihood-ratio systems. *Science & Justice, 51*(3), 91–98.
Neuhauser, S., & Simpson, A. P. (2007). Imitated or authentic? Listeners' judgements of foreign accents. In: *Proceedings of ICPhS 2007*, 1805–1808.
R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
Rhodes, R. (2017). Aging effects on voice features used in forensic speaker comparison. *International Journal of Speech, Language and the Law*, 24(2), 177–199.
Rogers, H. (1998). Foreign accent in voice discrimination: A case study. *Forensic Linguistics, 5*(2), 203–208.
Ross, S., Earnshaw, K., & Gold, E. (2019). A cautionary tale for phonetic analysis: the variability of speech between and within recording sessions. In: *Proceedings of ICPhS 2019*, 3090–3094.
Růžičková, A., & Skarnitzl, R. (2017). Voice disguise strategies in Czech male speakers. *Acta Universitatis Carolinae – Philologica 3, Phonetica Pragensia XIV*, 19–34.
Scherer, K. R. (2019). Acoustic patterning of emotion vocalization. In: Frühholz, S., & Belin, P. (Eds.), *Oxford handbook of voice perception* (pp. 61–91). Oxford University Press.
Schiller, N. O., Köster, O., & Duckworth, M. (1997). The effect of removing linguistic information upon identifying speakers of a foreign language. *International Journal of Speech, Language and the Law, 4*(1), 1–17.
Shriberg, E., & Scheffer, N. (2009). Does session variability compensation in speaker recognition model intrinsic variation under mismatched conditions? In: *Proceedings of Interspeech 2009*, 1551–1554.
Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. (2022). afex: Analysis of Factorial Experiments. R package version 1.1-1, <https://cran.r-project.org/package=afex>.
Skarnitzl, R., Asiaee, M., & Nourbakhsh, M. (2019). Tuning the performance of automatic speaker recognition in different conditions: Effects of language and simulated voice disguise. *International Journal of Speech, Language and the Law, 26*(2), 209–229.
Stoet, G. (2010). PsyToolkit – A software package for programming psychological experiments using Linux. *Behavior Research Methods*, 42(4), 1096–1104.
Stoet, G. (2017). PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology*, 44(1), 24–31.
Šturm, P., Skarnitzl, R., & Nechanský, T. (2021). Prosodic accommodation in face-to-face and telephone dialogues. In: *Proceedings of Interspeech 2021*, 1444–1448.

Sullivan, K., & Schlichting, F. (2000). Speaker discrimination in a foreign language: First language environment, second language learners. *Forensic Linguistics, 7*(1), 95–111.
Torstensson, N., Eriksson, E. J., & Sullivan, K. P. H. (2004). Mimicked accents – Do speakers have similar cognitive prototypes? In: *Proceedings of SST2004: the 10th Australian international conference on speech science and technology*, 271–276.

Tomáš Nechanský
Institute of Phonetics
Faculty of Arts, Charles University
Prague, Czech Republic
E-mail: tomas.nechansky@seznam.cz

Tomáš Bořil
Institute of Phonetics
Faculty of Arts, Charles University
Prague, Czech Republic
E-mail: tomas.boril@ff.cuni.cz

Alžběta Houzar
Institute of Phonetics
Faculty of Arts, Charles University
Prague, Czech Republic
E-mail: alzbeta.houzar@ff.cuni.cz

Radek Skarnitzl
Institute of Phonetics
Faculty of Arts, Charles University
Prague, Czech Republic
E-mail: radek.skarnitzl@ff.cuni.cz

2022 ACTA UNIVERSITATIS CAROLINAE PAG. 23–33
PHILOLOGICA 1 / PHONETICA PRAGENSIA

BILINGUAL ACOUSTIC VOICE VARIATION: THE CASE OF SORANI KURDISH-PERSIAN SPEAKERS

MARAL ASIAEE, HOMA ASADI

ABSTRACT

Many individuals around the world speak two or more than two languages. This phenomenon adds a fascinating dimension of variability to speech, both in perception and production. But do bilinguals change their voice when they switch from one language to the other? It is typically assumed that while some aspects of the speech signal vary for linguistic reasons, some indexical features remain unchanged across languages. Yet little is known about the influence of language on within- and between-speaker vocal variability. The present study investigated how acoustic parameters of voice quality are structured in two languages of a bilingual speaker and to what extent such features may vary between bilingual speakers. For this purpose, speech samples of 10 simultaneous Sorani Kurdish-Persian bilingual speakers were acoustically analyzed. Following a psychoacoustic model proposed by Kreiman (2014) and using a series of principal component analyses, we found that Sorani Kurdish-Persian bilingual speakers followed a similar acoustic pattern in their two different languages, suggesting that each speaker has a unique voice but uses the same voice parameters when switching from one language to the other.

Keywords: voice quality, bilingual speakers, Persian, Sorani Kurdish, principal component analysis

1. Introduction

Laver (1980) described voice quality as the "characteristic auditory coloring of an individual speaker's voice". Abercrombie (1967: 91) defined voice quality as "those characteristics which are present more or less all the time that a person is talking: it is a quasi-permanent quality running through all the sound that issues from his mouth".

While the anatomical and physiological characteristics of an individual's vocal apparatus play a role in the quality of the voice, some characteristics are shared amongst the members of the same linguistic community, i.e., speakers of a certain community have acquired the shared features to mark their social or regional membership in a group (Esling et al., 2019). This includes speakers of the same language, whose articulators, Esling (2000) believes, are physiologically trained to operate based on the phonetic constituents of that particular language. Honikman (1964) as cited in Esling (2000) proposed that activation of articulatory postures and patterns is a function of the language being

https://doi.org/10.14712/24646830.2022.26

spoken. Thus, one may assume that a switch from one language to another by bilinguals entails a variation in voice quality. Therefore, the present study seeks to find out whether the voice quality of the bilinguals varies across the two languages they speak.

Some researchers have investigated, with inconclusive results, what aspects of voice change and what aspects are robust against change across different languages. Amongst these surveys, F0 was the most studied voice quality feature. In the case of Cantonese-English bilinguals, Altenberg and Ferrand (2006) found no significant difference in F0 while the subjects were speaking either English or Cantonese. However, Ng et al. (2010) reported a correlation between F0 and the language being spoken, and higher F0 values were reported by Ng et al. (2012)including fundamental frequency (F0 when women were speaking English. Engelbert (2014) compared bilingual Brazilian's production of English and Portuguese. She found a significant difference in LTAS, F0, H1-H2, and harmonics-to-noise ratio (HNR) in the two languages. Lee and Sidtis (2017) did research on Korean-English and Mandarin-English speakers. Their results indicated that bilingual speakers in both language groups exhibited different voice patterns depending on the language. Johnson et al. (2020) investigated the degree to which the voice quality of bilingual speakers changes across two languages, namely Cantonese and English. They extracted and measured F0, F1-F4, the corrected versions of harmonic spectral slopes (i.e., H1*–H2*, H2*–H4* respectively), the corrected version of amplitude difference between the fourth harmonic and the harmonic closest to 2000 Hz (i.e., H4*–H2kHz*), the corrected amplitude difference between the harmonics closest to 2000 Hz and 5000 Hz (i.e., H2kHz*–H5kHz*), cepstral peak prominence (CPP), energy, and subharmonics-harmonics amplitude ratio (SHR), using VoiceSauce (Shue et al., 2009). They found that the majority of speakers have the same voice across the two languages. In a study done by Cheng (2020). The f0 level was higher for Korean\nthan English, regardless of gender, age, or generational status (early\nand late bilinguals did not differ, F0 was found to change in Korean-English bilinguals across the two languages.

As noted above, no consensus was achieved on whether bilinguals use the same voice in the two languages. For some speakers, no change was observed in their voice, while others change their voice quite substantially across the languages. Therefore, the main goal of the present research is to investigate if the voice quality changes in the case of Kurdish-Persian bilinguals.

2. Method

The present study is a pilot study of a project which is going to be done on a larger corpus of bilinguals, speaking different languages.

2.1 Data

Speech samples from 10 simultaneous male bilinguals of Kurdish-Persian were recorded. All participants were educated and spoke a Sorani variety of Kurdish. Their age range was between 25–39 (Mean= 34, SD= ±4.02). All speakers were asked to read "the North Wind and the Sun" once in Persian and once in Kurdish at their comfortable pitch and loudness and with their normal speaking rate in two different recording sessions.

The audio recordings were compiled using ZOOM H5 hand-held recorder that was set on 44.1 kHz and 16-bit resolution. The recorder was held 20 cm away from the speaker's mouth at a 45° angle. All recordings were done in a quiet room with no background noise.

2.1.1 Persian and Sorani Kurdish speech sounds

Both Persian and Kurdish belong to the Iranian branch of Indo-Iranian languages which itself is a branch of the Indo-European language family.

Persian is an aspiration language with 6 monophthongs (/i, e, a, ɑ, o, u/) and 23 consonants. While some scholars argue that Persian has six diphthongs (/ei/,/ai/, /ɑi/, /ui/,/oi/, /ou/), others believe that these are sequences of a vowel and a semi vowel. The syllable structure in Persian is CV(C)(C).

Kurdish belongs to the northern branch of western Iranian languages. The language itself is stratified into 3 different categories: Northern, Central, and Southern. Sorani is one of the central Kurdish varieties. It has 8 monophthongs /i, e, æ, ə, u, ʊ, o, ɑ/, 7 diphthongs, and 29 consonants. The syllable structure in Sorani Kurdish is considered to be the same as Persian, i.e., CV(C)(C).

2.2 Pre-processing the speech samples

Before carrying out the acoustic measurements of voice quality parameters, all voiced segments (vowels and consonants) of the signals were extracted using the command (Extract voiced and unvoiced) in Praat Vocal Toolkit (Corretge, 2022) which is a free plugin for Praat (Boersma & Weenink, 2022) with automated scripts for voice processing. Only the voiced parts were saved and used for further analysis. All voice quality measurements were done using the VoiceSauce (Shue et al., 2009) however, the default was changed to 5 ms intervals.

2.3 Acoustic parameters

Voice-quality-related acoustic parameters were selected based on the "psychoacoustic model of voice quality" proposed by Kreiman et al. (2014). According to Kreiman et al. (2014), only the "necessary" and "sufficient" parameters to model the voice quality are included in the model. The model components were originally stratified into four different categories, including "time-varying source characteristics", "vocal tract transfer function", "harmonic source spectral shape" and "inharmonic source excitation".

The parameters used in the present study and the original model are presented based on the category they belonged to.

F0: a parameter that depicts "the time-varying source characteristic" (Kreiman et al., 2014) and is a perceptual correlate of the pitch.

The first four formant frequencies (F1, F2, F3, F4): the first four formant frequencies are associated with the transfer function of the vocal tract (Kreiman et al., 2014). The first three formants are commonly employed when discussing the linguistic variations

in different languages and the fourth formant is mostly referred to as a speaker-specific parameter (Johnson et al., 2020)

H1-H2*1, H2*-H4*, H4*–H2kHz*, and H2kHz*–H5kHz:* all these parameters are associated with the spectral shape of the harmonic source. H1*-H2* denotes the difference between the amplitude of the first and the second harmonics and gauges the harmonic slope which is indicative of the phonation type. H2*-H4* is the relative amplitude of the second and the fourth harmonic in a higher frequency band. H4*–H2kHz*is the difference between the amplitude of the fourth harmonic and the harmonic nearest to the 2000 Hz in frequency. This parameter measures the spectral slope of the harmonic in a higher frequency band. H2kHz*–H5kHz is the amplitude difference between the closest harmonics to the 2000 Hz and 5000 Hz. This parameter is related to the spectral slope of harmonic independent of F0 (Johnson et al., 2020).

Cepstral peak prominence (CPP): corresponds to the ratio of harmonic energy to spectral noise. It is correlated with the degree of the regularity and periodicity of the voice signal (Hillenbrand, 2011)

Apart from the parameters in the original model, several other parameters were added to it including formant dispersion (FD), energy, and subharmonics-harmonics ration (SHR) (Y. Lee et al., 2019). FD is the "averaged distance between successive formant frequencies" and is believed to be associated with vocal tract length (Fitch, Energy refers to "the Root Mean Square (RMS) energy, calculated at every frame over a variable window equal to five-pitch pulses"(Shue et al., 2009). SHR quantifies the amplitude ratio of subharmonics to harmonics and is related to period-doubling. The spectral noise is characterized by these two parameters in addition to the CPP. The last acoustic measure that was added to the original model was the *moving coefficients of variations* (moving $\text{CoV}=\frac{moving\ standard\ deviation\ (\sigma)}{moving\ mean\ (\mu)}$) to capture the dynamic variations of voice quality since it is believed that listeners do not only rely on the absolute values of different measures to discriminate between speakers (Y. Lee & Kreiman, 2019). Table 1 represents the parameters and their corresponding categories.

Table 1. Acoustic measures with their corresponding categories

Category	Parameter
F0	F0
Formants	F1, F2, F3, F4, FD
Harmonic source spectral shape	H1*–H2*, H2*–H4*,H4*–H2kHz*, H2kHz*–H5kHz
Inharmonic source/spectral noise	CPP, Energy, SHR
Variability	Coefficients of variation for all acoustic measures

[1] The asterisk sign(*) accompanying the harmonics signifies that the parameters are corrected for the effect of formants on harmonic amplitudes (Iseli & Alwan, 2004; Lee et al., 2019)

2.4 Post-processing the acoustic parameters

After running the VoiceSauce, observations with erroneous values (e.g, impossible 0 value for F0) were removed from the data set. Per speakers, values of each parameter were then normalized regarding the minimum and maximum value of that parameter in the whole data set in each language. The final values of each parameter ranged from 0 to 1 after normalization. The moving coefficient of variation for each parameter was calculated using a 50 ms window (10 observations). In total, 102114 data frames from Kurdish samples and 104608 data frames from Persian samples were obtained.

2.5 Statistical analysis

The method used in this survey is adapted from the works of Johnson et al. (2020), Lee et al. (2019), and Lee & Kreiman (2022) on analyzing the voice-quality-related parameters. All analyses were carried out using R version 4.0.4 (R Core Team, 2021)

An independent sample t-test was conducted to find out whether acoustic measurements remain stable or vary across the languages in general and in each individual in particular.

Then, to extract the internal structure of the data in the present study we performed Principal component analysis (PCA). PCA is a method used to reduce the dimensionality of large data sets while at the same time making the interpretation of the results easier. In PCA variables that are on the one hand correlated with one another and on the other hand independent of other groups of variables, are categorized into one component. Breaking down the data sets into different components enables the researcher to better identify and explain the internal structure of the data- and thus find the similarities and differences in the voices of bilinguals. In PCA, since some correlation was expected between the measured parameters, oblique rotation was implemented to simplify the structure of the data (Johnson et al., 2020; Y. Lee et al., 2019; Y. Lee & Kreiman, 2022). Only components with eigenvalues greater than 1 were included to ensure the interpretability of variances in the data (Kaiser, 1960). The loadings (weight) threshold for the parameters to be included in a component is equal to or higher than 0.32.

First, the common voice space for each language was designated by performing PCA in the whole Persian and Sorani Kurdish data sets. By doing so, the difference between the internal structure of both languages was captured.

Second, PCA was separately conducted for each speaker in each language using all 26 acoustic measurements obtained from the speech samples of that individual (13 variables + 13 CoVs for each variable), i.e., we had two PCAs per speaker (one in Persian, one in Sorani Kurdish). Then, the cumulative number of times each parameter appeared in each component was calculated by counting the times a particular parameter (e.g., F0) appeared in each component (the data is comprised of speech samples from 10 speakers in each language, therefore, no matter which component a particular parameter appears in, the cumulative number, in the end, would be 10). In this way, differences between the individuals in each language will be accounted for (individual voice space). Then the most prominent parameter in each category was determined. This was done in order to extract the general voice space within the individuals.

3. Results and Discussion

Results from t-test analysis showed that while all F0, formants, source spectral shape, and spectral noise parameters remained stable across Persian and Sorani Kurdish, almost all CoVs (except CoV F1 and CoV SHR) varied significantly. The effect size of the difference between the parameters across languages, however, was trivial. Detailed results for each variable parameter are presented in Table 2.

Table 2. Results obtained from independent sample t-test run on the whole Persian and Sorani Kurdish data set

Parameter	Results	Cohen's d	Median	
			Persian	Kurdish
CoV F0	$t(206719) = -70.240$, $p < 0.05$	0.309	0.088	0.076
CoV F2	$t(206719) = 2.906$, $p < 0.05$	0.013	0.199	0.204
CoV F3	$t(206719) = -4.00$, $p < 0.05$	0.018	0.132	0.126
CoV F4	$t(206719) = 14.150$, $p < 0.05$	0.062	0.219	0.224
CoV FD	$t(206719) = 14.178$, $p < 0.05$	0.062	0.219	0.224
CoV H1*–H2*	$t(206719) = -2.833$, $p < 0.05$	0.012	0.128	0.125
CoV H2*–H4*	$t(206719) = 8.320$, $p < 0.05$	0.037	0.185	0.187
CoV H4*–H2kHz*	$t(206719) = 11.069$, $p < 0.05$	0.049	0.186	0.183
CoV H2kHz*–H5kHz	$t(206719) = -3.539$, $p < 0.05$	0.011	0.149	0.146
CoV CPP	$t(206719) = 14.703$, $p < 0.05$	0.065	0.231	0.233
CoV Energy	$t(206719) = -34.257$, $p < 0.05$	0.151	0.057	0.054

Since some differences were observed between the acoustic parameters of voice, PCA was conducted to extract the common voice space of each language and find out how similar and/or different acoustic voice spaces are structured across Persian and Sorani Kurdish.

PCA resulted in 11 components for each language which cumulatively accounted for 68.9% and 70.5% of variances in Persian and Sorani Kurdish respectively. Analyzing the parameters in each component revealed that there is a similarity in the occurrence of the parameters in each component, specifically those parameters that did not exhibit significant variation in the t-test analysis. The internal structure of each component is represented in Figure 1.

As can be observed in Fig. 1, those parameters that did not vary across the languages either appear in the same components (PC01, PC02, PC05, PC11) in both languages or they appear in combination with the same other parameters (F0 and Energy in PC09 and PC06, F1 and CoV F1 in PC10 and PC09 in Persian and Sorani Kurdish respectively). The four components that were completely similar across the languages, accounted for 32.1% of the variability in Persian and 32.9% in Kurdish. The difference between the two

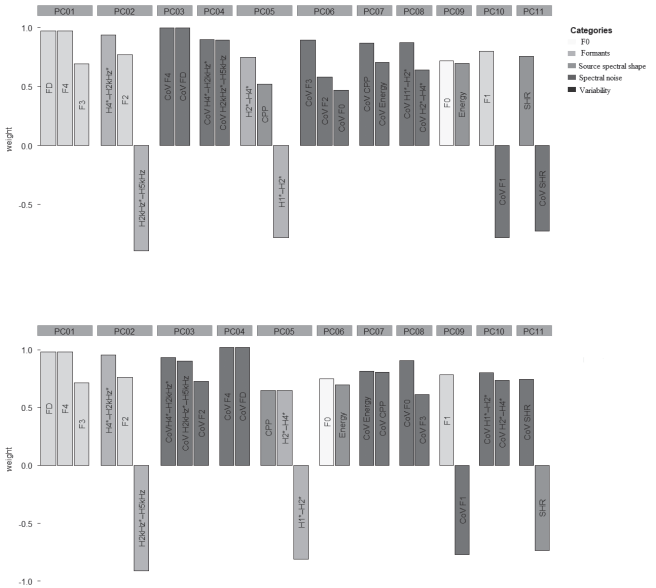

Figure 1. Bar plots of acoustic parameters in all PCs for Persian speakers (Top panel) and Sorani Kurdish speakers (bottom panel). Parameters in each PC are ordered from the highest absolute value of rotated component loading (weight) to the lowest. The hue of each bar delineates the category of the parameter.

languages is mostly observed where coefficients of variations (CoVs) emerged in the components. These parameters are the ones that differed significantly across the languages, therefore, variation in them was expected.

Based on the results obtained, acoustic measures of formant frequencies (FD, F4, F3) were dominant in the first component for both languages, accounting for 10.9% of the variance in the Persian data set and 11.7% in Sorani Kurdish. The second component was predominantly occupied by spectral shape measures (H4*–H2kHz* and H2kHz*–H5kHz) and formant frequency measures (F2), representing 10.2% of the variance in Persian and 11% in Sorani Kurdish.

The third component in Persian consists of coefficients of variation for F4 and FD, while these measures appear in the fourth component for Sorani Kurdish. The third component in Sorani Kurdish was strongly based on the coefficient of variation for the spectral shape measures and CoV F2.

Since some differences were observed between the acoustic parameters of voice across languages, a student's t-test was employed to find out how variable these parameters were across each individual's voice. The same results as the cross-language analysis were obtained for the variation of the acoustic parameters in each individual albeit with a difference in their effect size and inclusion of CoV F1 and CoV SHR in the results. The number of speakers whose acoustic voice quality parameters vary significantly is reported in Table 3.

Table 3. Number of Cohen's d for cross-linguistic comparisons in parameters that differed significantly in each individual

Parameter	Number	Cohen's d			
		Trivial 0–0.2	Small 0.2–0.5	Medium 0.5–0.8	Large > 0.8
CoV F0	9/10	5	2	1	1
CoV F1	1/10	1	–	–	–
CoV F2	10/10	5	5	–	–
CoV F3	8/10	3	4	1	–
CoV F4	10/10	4	3	3	–
CoV FD	10/10	4	3	3	–
CoV H1*–H2*	10/10	4	4	2	–
CoV H2*–H4*	8/10	4	4	–	–
CoV H4*–H2kHz*	10/10	4	6	–	–
CoV H2kHz*–H5kHz	8/10	3	5	–	–
CoV CPP	9/10	3	5	–	1
CoV Energy	10/10	5	3	2	–
CoV SHR	*5/10*	5	–	–	–

Observing the difference in some parameters in each individual, we performed separate PCAs for all the speakers and then determined the most prominent parameter in each component based on the results obtained from each speaker. In this way, while a common pattern amongst speakers was identified, the speaker-specific patterns were considered as well (see 2.5 for a detailed explanation of the method). Figure 2 delineates the internal structure of PCA results.

Since Figure 2 represents the general voice space within the individuals, variations in the occurrence of parameters in the components were expected. As is evident from Fig. 2, the most prominent category of parameters in components 1 to 3 are the formants and their CoVs counterparts in both Persian and Sorai Kurdish with the addition of source spectral shape parameters in the Persian data set. This is the same pattern that has emerged as the common voice space in Persian and Sorani Kurdish, albeit with different ordering and weight of the parameters in each component. Spectral slope in the higher frequencies (H4*–H2kHz* and H2kHz*–H5kHz) appeared along with F2 in Persian which is similar to their occurrence in the common voice space of Persian and Sorani Kurdish. The first three components in Persian and Kurdish accounted for 28.39% and 28.47% of variances respectively.

Like the common voice in Persian and Sorani Kurdish, F0 did not emerge in lower-order components, but when it did, it was accompanied by Energy. F1 and SHR with their CoV counterparts emerged in the same components in both languages.

Overall, the results obtained here were consistent with the results in Johnson et al. (2020) which showed that the acoustic patterns of voice were similarly structured across

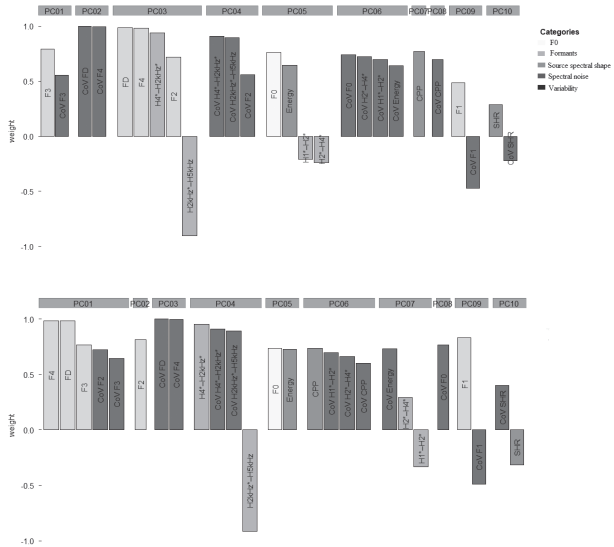

Figure 2. Bar plots of acoustic parameters emerging in 10 PCs for Persian speakers (Top panel) and Sorani Kurdish speakers (bottom panel), depicting the general voice space within the individuals. Parameters in each PC are ordered from the highest absolute value of rotated component loading (weight) to the lowest. The hue of each bar delineates the category of the parameter

the two languages of bilingual speakers. This delineates that bilingual speakers have the same voice when they switch from one language to another.

3. Conclusion

The present research studied how acoustic voice spaces vary individually and commonly across two languages of bilingual speakers. Results revealed that acoustic voice variations are similarly structured in the different languages of the speakers. While some acoustic voice space is shared amongst speakers across their two languages, there are also speaker-specific patterns within individual speakers, suggesting that each speaker has his/her own unique pattern as well. This means that speakers vary in the extent of acoustic voice quality structures between themselves. However, they showed a substantial similarity toward themselves. Despite the phonetic differences in the speech sound patterns of Sorani Kurdish and Persian, the variation in acoustic voice quality revealed a similar pattern in the lower dimensional structures.

Acknowledgment

This work was supported by Iran National Science Foundation (INSF) Grant No. 99029580.

REFERENCES

Abercrombie, D. (1967). *Elements of General Phonetics*. Edinburgh University Press.

Altenberg, E. P., & Ferrand, C. T. (2006). Fundamental Frequency in Monolingual English, Bilingual English / Russian, and Bilingual English / Cantonese Young Adult Women. *Journal of Voice*, 20(1), 86–96. https://doi.org/10.1016/j.jvoice.2005.01.005

Boersma, P., & Weenink, D. (2022). *Praat: Doing phonetics by computer* (6.2.09). https://www.fon.hum.uva.nl/praat/

Cheng, A. (2020). Cross-linguistic f 0 differences in bilingual speakers of English and Korean. *The Journal of the Acoustical Society of America*, 147(2), EL67–EL73. https://doi.org/10.1121/10.0000498

Corretge, R. (2022). *Praat Vocal Toolkit*. http://www.praatvocaltoolkit.com

Engelbert, A. P. P. F. (2014). Cross-Linguistic Effects on Voice Quality : A Study on Brazilians ' Production of Portuguese and English. Proceedings of the International Symposium on the Acquisition of Second Language Speech. *Concordia Working Papers in Applied Linguistics* 5, 157–170.

Esling, J. H. (2000). Crosslinguistic aspects of voice quality. In R. D. Kent & M. J. Ball (Eds.), *Voice Quality Measurement* (pp. 25–35). Singular Publishing Group.

Esling, J. H., Moisik, S. R., Benner, A., & Crevier-Buchman, L. (2019). Voice and Voice Quality. In *Voice Quality: The Laryngeal Articulator Model* (pp. 1–36). Cambridge University Press. https://doi.org/10.1017/9781108696555.001

Fitch, W. T. (1997). Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques. *Journal of Acoustical Society of America*, 102(2), 1213–1222. https://doi.org/10.1121/1.421048

Hillenbrand, J. M. (2011). *Acoustic Analysis of Voice: A Tutorial. Perspectives on Speech Science and Orofacial Disorders*, 21(2), 31. https://doi.org/10.1044/ssod21.2.31

Honikman, B. (1964). Articulatory settings. In D. Abercrombie, D. B. Fry, P. A. D. MacCarthy, N. C. Scott, & J. L. M. Trim (Eds.), *In honour of Daniel Jones: Papers contributed on the occassions of his eightieth birthday 12 September 1961* (pp. 73–84).

Iseli, M., & Alwan, A. (2004). An improved correction formula for the estimation of harmonic magnitudes and its application to open quotient estimation. *Proceedings of IEEE ICASSP*, 1, 10–13.

Johnson, K. A., Babel, M., & Fuhrman, R. A. (2020). *Bilingual acoustic voice variation is similarly structured across languages. Proceedings of the Annual Conference of the International Speech Communication Association*, INTERSPEECH, 2020-Octob, 2387–2391. https://doi.org/10.21437/Interspeech.2020-3095

Kaiser, H. F. (1960). *The Application of Electronic Computers to Factor Analysis. Educational and Psychological Measurement*, 20(1), 141–151. https://doi.org/10.1177/001316446002000116

Kreiman, J., Gerratt, B. R., Garellek, M., Samlan, R., & Zhang, Z. (2014). Toward a unified theory of voice production and perception. *Loquens*, 1(1), e009. https://doi.org/10.3989/loquens.2014.009

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

Lee, B., & Sidtis, D. V. L. (2017). The bilingual voice: Vocal characteristics when speaking two languages across speech tasks. *Speech, Language and Hearing*, 20(3), 174–185. https://doi.org/10.1080/2050571X.2016.1273572

Lee, Y., Keating, P., & Kreiman, O. (2019). Acoustic voice variation within and between speakers. *The Journal of the Acoustical Society of America*, 146(3), 1568–1579. https://doi.org/10.1121/1.5125134

Lee, Y., & Kreiman, J. (2019). *Within- and between-speaker acoustic variability : Spontaneous versus read speech. October, 1–2*. https://doi.org/10.1121/1.5137431

Lee, Y., & Kreiman, J. (2022). Acoustic voice variation in spontaneous speech. *The Journal of the Acoustical Society of America*, 151, 3462–3472. https://doi.org/10.1121/10.0011471

Ng, M. L., Chen, Y., & Chan, E. Y. K. (2012). Differences in vocal characteristics between Cantonese and English produced by proficient Cantonese-English bilingual speakers – A long-term average spectral analysis. *Journal of Voice*, 26(4), e171–e176. https://doi.org/10.1016/j.jvoice.2011.07.013

Ng, M. L., Hsueh, G., & Leung, C. S. A. M. (2010). Voice pitch characteristics of Cantonese and English produced by Cantonese- English bilingual children Voice pitch characteristics of Cantonese and

English produced by Cantonese-English bilingual children. *International Journal of Speech-Language Pathology*, 12(3), 230–236. https://doi.org/10.3109/17549501003721080
R Core Team. (2021). R: *A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*. https://www.r-project.org/
Shue, Y.-L., Keating, P., & Vicenik, C. (2009). VOICESAUCE: A program for voice analysis. *The Journal of the Acoustical Society of America*, 126(4), 2221. https://doi.org/10.1121/1.3248865

Maral Asiaee
Department of Linguistics,
Alzahra University
Tehran, Iran
E-mail: m.asiaee@alzahra.ac.ir

Homa Asadi
Department of Linguistics,
University of Isfahan
Isfahan, Iran
E-mail: h.asadi@fgn.ui.ac.ir

2022 ACTA UNIVERSITATIS CAROLINAE PAG. 35–50
PHILOLOGICA 1 / PHONETICA PRAGENSIA

WORD-BOUNDARY GLOTTALIZATION AND ITS COGNITIVE ASPECTS: A REACTION-TIME STUDY

MICHAELA SVOBODA, PAVEL ŠTURM

ABSTRACT

This paper deals with the perceptual impact of using word-initial glottalization as a boundary signal. Research into glottalization is discussed before presenting a reaction time (RT) experiment based on a word-monitoring paradigm that aims to investigate the cognitive effect of the presence of glottalization in native Czech listeners' perception. Natural and acoustically manipulated sentences selected from a spontaneous political debate were used as stimuli. The low level of control of the material makes the design relatively innovative in comparison with similar previous studies. Fifty listeners were instructed to press a button as soon as they heard a specified target word embedded in a sentence, where a preceding carrier word included glottalization or linking. The results support the hypothesis that the presence of glottalization facilitates the processing of subsequent words, but the temporal scope of the effect varied. The experiment also raises new methodological issues and offers suggestions for further RT research.

Key words: cognition, Czech, glottalization, perception, reaction time

1. Form and function of glottalization

This paper examines 'glottalization', a specific phonetic event originating in the larynx that has been given a variety of labels. Since terminological inconsistencies and idiosyncratic interpretations can generate only confusion in the field, we shall start with defining and clarifying the relevant terms, for which differentiating between form and function will be a key prerequisite.

In the narrow, articulatory sense, the GLOTTAL/LARYNGEAL STOP (GLOTTAL/LARYNGEAL PLOSIVE) is a non-continuant obstruent articulation at the glottis, i.e., a complete closure at the adducted vocal folds. The air is compressed in front of the obstruction until the closure is released, resulting in a short burst or 'plosion'. In this perspective, the glottal stop is just one of the items in the IPA row of plosives, transcribed as [ʔ]. Acoustically, glottal plosives are not all identical, as they differ in the shape of the waveform (see Skarnitzl, 2004 for details and images). What is common to them is a percept of a sudden change – cessation to silence or rise from silence (Cruttenden, 2014: 182).

https://doi.org/10.14712/24646830.2022.27

However, in identical positions, in free variation with the glottal plosives, other speech sounds may occur. Although they are quite different from plosives acoustically, they are quite similar perceptually (namely, a break between the two neighbouring sounds is perceived). The glottal plosive may be seen as an extreme position on the phonation continuum (Gordon & Ladefoged, 2001: 384), in which the non-modal glottal phenomena between modal phonation and glottal closure are perceptually equivalent. In fact, in genuine spoken interactions, complete glottal closures are less common than various lenited variants, including creaky voice or laryngealization characterized chiefly by irregularity in vibration and its lower frequency (Ashby & Przedlacka, 2011, 2014, for English). Several publications deal with the acoustics of these glottal events (Docherty & Foulkes, 1999; Redi & Shattuck-Hufnagel, 2001; Skarnitzl 2004; Keating, Garellek & Kreiman, 2015).

All the manifestations mentioned above can be subsumed under **one common term**, namely glottalization. Any such glottal activity – glottal plosive, creaky voice – would be perceived as a single phenomenon – glottalization. Sometimes, the term glottal stop is used in this broader sense as well. We will be referring to 'glottalization' in the rest of the paper. However, not every occurrence of creaky phonation (laryngealization) is an instance of glottalization. For example, creaky phonation may be associated with ends of prosodic phrases (Local, Wells & Sebba, 1985; Ogden, 2001), be a habitual (sociolinguistic) feature of one's voice (Yuasa, 2010; Wolk, Abdelli-Beruh & Slavin, 2012) or can signal a pragmatic or affective meaning (Gobl & Ní Chasaide, 2003). This type of creak seems to be prosodically grounded compared to the segmentally defined creak in glottalization.

Finally, it is important to establish the **function of glottalization** in the given language. As a functionally defined unit, glottalization can be employed in two main ways. It can be used (1) contrastively, where [ʔ] is in opposition to realizations of other consonant phonemes. In languages such as Hawaiian or Arabic, it is the primary allophone of the phoneme (Ladefoged & Maddieson, 1996; Maddieson, 1984), while in languages such as English or the Philippine language Ilokano, it is a regional or stylistic variant (Cruttenden, 2014; Olaya, 1967). Specifically, the oral plosive [t] (*cat, later*) in one accent of English may correspond to [ʔ] (glottalization) in other accents, or it may be conditioned by the environment or speaking style within one accent (see e.g., Fabricius, 2002; Gavaldà, 2016). Moreover, glottalization is often used as (2) a boundary signal of a lexical/grammatical unit, i.e., it has a demarcative function. Czech or Polish are typical examples of languages where the presence of glottalization cues morphosyntactic word boundaries when the word starts with a vowel ([ʔ]*akustika*, 'acoustics'), and some internal morpheme boundaries (*na*[ʔ]*opak*, 'on the contrary'). The two functions are not incompatible, as some languages (e.g., English) have both contrastive and demarcative glottalization.

An anonymous reviewer pointed out another attested function, (3) hiatus breaking, in which sequences of two vowels are split by glottalization as a sort of syllable reorganization or onset formation (for instance in German *Theater* [teˈʔaːtɐ], 'theatre' or Czech *teoreticky* [ˈtɛʔorɛtɪt͡skɪ] 'theoretically'). However, the use of such forms needs to be examined empirically.

2. Use of boundary glottalization in languages

This paper deals with glottalization in its demarcative function, i.e., as a boundary cue. Available sources suggest that its fulfilment is not a binary matter; rather, there is a continuum between the tendency to glottalize initial vowels and the opposite tendency to link initial vowels to the previous words. The studies presented below offer evidence of substantial difficulty in trying to establish an 'inherent' glottalization rate for each language, as it seems that **context (speaking situation)** is a key factor. Therefore, it is imperative to consider the type of the examined material, especially its position on the scale of spontaneity or level of control (Wagner, Trouvain & Zimmerer, 2015). In cross-linguistic comparison, one cannot confront for instance the material of read speech with a spontaneous dialogue.

Nevertheless, owing to considerable research interest in the topic, a certain picture on the use of glottalization emerges at least for some languages. For instance, German is often characterized as a language that employs glottalization quite extensively. Kohler (1994) measured the rate of glottalization in read speech from the Phonetic Data Bank of German, yielding 79% of words with initial vowels being glottalized. This study also suggested that a higher glottalization rate can be expected in stressed syllables, which was confirmed in subsequent research (Pompino-Marschall & Żygis, 2010; Malisz, Żygis & Pompino-Marschall, 2013). Moreover, these later studies were of somewhat lower level of control, represented by prepared speech, and established, respectively, 50–70% and 63% of glottalized initial vowels. Other contributing factors were increased speech tempo, leading to higher glottalization rates, or the semantic status of the word, with lexical words being glottalized more often than grammatical words.

Moving to the opposite tendency, English is a language that uses linking rather than boundary glottalization (Cruttenden, 2014). However, this does not mean that we should expect no glottalization at word edges at all (see previous section for contrastive glottalization). For instance, Dilley, Shattuck-Hufnagel and Ostendorf (1996) showed that news-reading was associated with substantial variability across speakers, ranging from 13% to 44%. Another finding was the contribution of prosodic factors, including lexical stress (as in German) or the strength of the prosodic boundary (stronger boundaries tended to involve more glottalization). Many Romance languages – French, Spanish, Italian, Portuguese – behave similarly to English in this respect (Skákal, 2013; Di Napoli, 2015; Skarnitzl, Čermák, Šturm, Obstová & Hricsina, 2021). The main strategy in these languages is to link words smoothly without glottalization (thus *por‿otro, que‿ahora* and not *por*[ʔ]*otro* etc.).

The research on Czech examined several factors and used material with various levels of control. Veroňková (2016) confirmed that the requirement to use glottalization after non-syllabic prepositions (Hála, 1967) was to a large extent obeyed in read speech. Volín (2012) brought a comparison of professional news-reading with semi-spontaneous dialogues. Glottalization rate was high in the former speaking style (97% by female newsreaders, 88% by their male colleagues) and much lower in the latter situation (men in spontaneous dialogues glottalized only 41% of the potential cases, women 65%). The results thus suggest an effect of material/context on the one hand, and gender on the other (compare similar effects of gender in Dilley et al., 1996;

Skákal, 2015; Gráczi & Markó, 2018; Kopečková, 2020). Further factors include the segmental and prosodic environment. Palková (2016), examining the read speech of students, compared V#V and C#V sequences in various positions within a prosodic phrase. Although the segmental environment did not play a significant role (similarly to Veroňková, 2016), strength of the prosodic boundary was a crucial factor (see also Dilley et al., 1996).

3. Perception of boundary glottalization

To fully understand a sound pattern such as glottalization, its perceptual impact must be thoroughly examined. The use of glottalization can be experimentally treated from several perspectives but for reasons of scope we will limit ourselves to discussing the effect of glottalization on the online processing of speech, namely on how words are recognized in an utterance.

A comparison of a glottalized sequence with a non-glottalized (linked) sequence is necessary. If glottalization plays an important role in processing the stream of speech, the former should have a processing advantage over the latter. Two situations should be distinguished. On the one hand, a vowel-initial word is cued with glottalization, thus comparing for instance *stál_a čekal* with *stál* [ʔ]*a čekal* ('He stood there and waited.'). On the other hand, two actual words differ solely in the presence of glottalization, such as *kočku* ('cat') × *k*[ʔ]*očku* ('to a loop'), *sokem* ('enemy') × *s*[ʔ]*okem* ('with an eye'), *kůlu* ('pole') × *k*[ʔ]*úlu* ('to a hive'), so that the use of glottalization would be crucial to understanding. However, it is remarkably difficult to find such oppositions. Moreover, the grammatical and pragmatic environment would make it even more difficult to provide a source of confusion as to which member of the pair is present. The context is sufficient to disambiguate the variants in a sentence.

However, it would be wrong to conclude that using or not using glottalization is entirely irrelevant to the online processing of speech. The brain operates as a predictive device that constantly lowers the amount of uncertainty about the upcoming events (Grossberg, 2003). As the signal unfolds, the brain matches the auditory encoding of the current acoustic signal to a set of stored representations. Lexical access activates a number of representations ('words') based on past experience (e.g., context, topic, a closed × open set of answers to a question) and on the auditory signal. For instance, hearing [kot] might activate the words *kotel*, *kotva*, *kotouč*, but also *k odkazu* [kotkazu], *k odpovědi* [kotpovjɛɟɪ] and so on. In contrast, hearing [kʔot] might skip the irrelevant words, as *kotel*, *kotva*, *kotouč* do not include the glottal stop in their representation (the speaker has never heard them with glottalization, so has not stored it as such), and only the words starting with /ot/ after the preposition *k* will be the basis of assembling the candidate list. Although the brain will eventually solve the segmentation problem in both cases, [kot] and [kʔot], not least because of the contextual disambiguation, it is safe to assume that **the presence of glottalization provides a processing benefit** and is cognitively less demanding than the non-glottalized variants.

To our knowledge, only three experiments have been designed to test this prediction explicitly (Bissiri, Lecumberri, Cooke & Volín, 2011; Volín, Uhrinová & Skarnitzl,

2012; Schwartz, Rojczyk & Balas, 2015). They all employ a word-monitoring paradigm, in which a subject's response to a specified target word is collected, measuring the reaction time (RT) of the response. The material was read English, but the listeners differed in their native languages and their typical glottalization rates. Bissiri et al. (2011) compared Czech, Spanish and native English listeners, Volín et al. (2012) Czech, Slovak and native English listeners, and Schwartz et al. (2015) used Polish listeners of varying English experience. The hypothesis was that the presence of glottalization would decrease the RT to that word compared to the linking condition. The two early studies in addition supposed that Czech listeners would benefit more from the presence of glottalization than English, Spanish or Slovak listeners, who use glottalization less extensively. Surprisingly, all groups were associated with decreased RTs in response to glottalization (despite some differences). The Polish experiments found an effect in the predicted direction, but it was small in size and also not much reliable.

However, the Czech studies differed from the Polish study in several important respects. Schwartz et al. (2015) included mostly predictable short sentences (*Bob ate the whole chicken*.), and the occurrence of glottalization or linking was always natural (produced by a real speaker in that context). In contrast, the material for Bissiri et al. (2011) and Volín et al. (2012) was news-reading, and the sentences started in medias res, with the beginning cut off so that the sentence would be unpredictable (*with ten men after the striker Thierry Henry*; the word *after* served as a target). Moreover, each sentence was presented in two conditions, with or without glottalization, which was done by means of manipulation of the speech signal (adding or removing glottalization), resulting in pairs of stimuli differing solely in that respect.

4. Experiment with RT measurement

The studies mentioned above examined the effect of glottalization in L2 English on non-native listeners. Whereas Czech listeners seem to benefit from the presence of initial glottalization in English, it is also necessary to determine the perceptual impact of boundary glottalization using native Czech material. The current experiment aims to fill that void. Our second goal is to move beyond scripted speech and use material with a lower level of control (Wagner et al., 2015), namely a political TV debate.

4.1 Method

4.1.1 Material

The recordings come from the political discussion TV programme *Nedělní partie*, in which a moderator (male, 31 years) hosts two guests (in our case, both male, 39 and 55 years) who present their views and argue with the moderator or with each other. The style is therefore semi-formal, the type of material can be labelled as non-scripted speech (Wagner et al., 2015) rather than truly spontaneous speech. The particular 50-minute episode was aired on TV Prima in the Czech Republic on 25/4/2010. The debate reflects some of the important political topics of the time.

The debate comprises 5841 orthographic words. Of these, 650 words (11%) begin with an initial vowel, which is typical for Czech initial syllables (12.2% according to Šturm & Bičan, 2021: 369). We conducted a simple acoustic analysis of the debate to determine the rate of boundary glottalization of the three speakers (word-internal glottalization was not examined). Tokens that were preceded by a pause or where several speakers spoke simultaneously were ignored. The host produced glottalization in 38%, and the two guests in 44% and 54% of the possible contexts. Such rates seem to be quite low, but it needs to be stressed that previous reports (e.g., Palková, 2016; Veroňková, 2016; partly Volín, 2012) were based on read material.

4.1.2 Stimuli for the perception test

Due to the low level of control of the material, we had no means of determining or shaping the sentences; it was only possible to select suitable items from the corpus. Such recordings (carrier + prompt embedded in sentences or their parts) had to be 'clean' (i.e., with no hesitations, interruptions, overlaps, or external noise), the site of glottalization should be suitable for unnoticed manipulation (see below), and the prompt should not have strong collocations with the preceding words, as this could weaken any perceptual effect.

The resulting set of 40 items (see the Appendix) is necessarily – but also intentionally – heterogenous. Twenty items occurred with glottalization, twenty without glottalization. Frequencies of the words, semantics, stress position, tempo or other factors were not controlled. This should not pose a problem, as we used planned paired comparisons between identical sentences where one version includes glottalization, while the other does not. Therefore, any such effect should be constant within each text. Importantly, the position of the prompt varied, and the beginning of the items does not correspond to sentence beginnings.

The manipulation of the carrier word was performed in Praat (Boersma & Weenink, 2021). The process differed depending on the sequence (C#V, V#V, types of V) and the type of manipulation (adding/removing glottalization). Adding glottalization was relatively easy: a glottal stop or its equivalent was copied from a suitable context from a different part of the debate, replacing the original linked context. Removing glottalization was more difficult. Again, a similar sequence was found in the debate – one without glottalization – which replaced the original glottalized token, or the glottalization interval was deleted. The neighbouring vowels usually had to be compensated in duration and adjusted (shortened or lengthened) manually or by PSOLA manipulations.

One of the key aspects that differentiates our experiment from those mentioned in Section 3 is the structure of the stimuli. In the previous setups, the target word or (i.e., the prompt, which is monitored for by the listeners) was at the same time a carrier word (i.e., the site of glottalization). The word *after* – with or without glottalization – was therefore printed on screen and reacted to in Volín et al.'s (2012) experiment. The current experiment involves a sequence of the manipulated carrier word followed by the target word (prompt), with varying distance of 1 to 7 syllables between the initial syllables. Therefore, the carrier is vowel-initial, but the target word does not start with a vowel (see Figure 1).

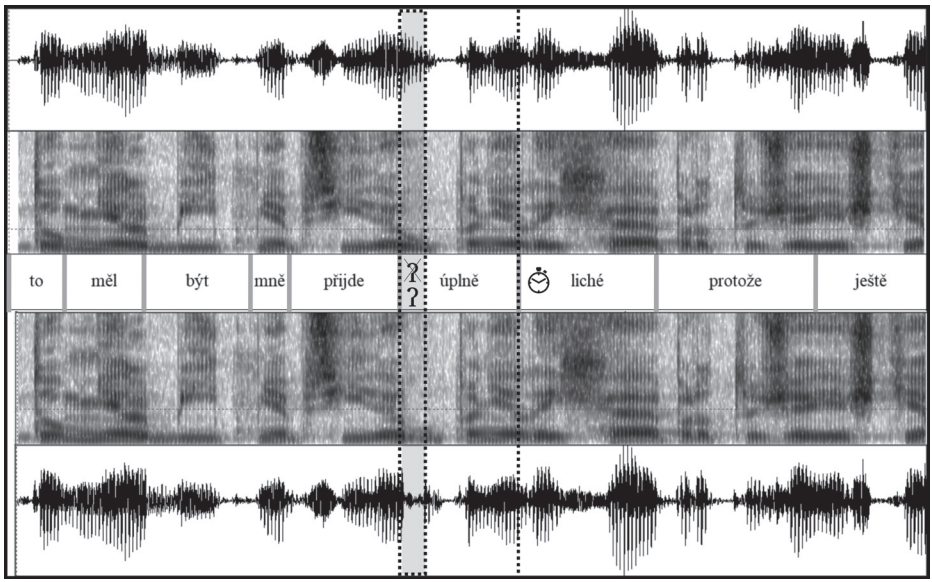

Figure 1. Example of a stimulus pair. The word *liché* is the prompt (target) displayed on the screen. The word *úplně* is the carrier – at its beginning, there is either linking (top) or glottalization (bottom). RT is measured from the start of the prompt (dotted line with a clock symbol). The prompt here is three syllables away from the glottalization site.

Such a change in the design of the experiment was done for the following reasons. First, it is not clear from which point RT should be measured in a vowel-initial word, as glottalization interferes with the temporal structure of the vowel. Having the onset of measurement on the following word, which is identical in both cases, solves this dilemma. Second, it is reasonable to assume that the effect of glottalization appears with a delay, i.e., the processing of the following word(s) would be affected. Finally, the modification also reflects our interest in knowing whether the distance between the carrier and prompt plays an important role. In some items, the glottalized syllable immediately precedes the prompt, in others it precedes it at a distance.

Two sets of 20 stimuli were originally produced with or without glottalization (Appendix). Two sets of 20 new stimuli were created by manipulation (adding or removing glottalization from the previous two sets). Therefore, 2 × 40 target items were used. In addition, 32 other items (almost 30% of the total number) served as fillers and distractors, in which the prompt occurred very early or late in the sentence, or it did not occur at all. Inclusion of such trials makes the task of recognizing a word more interesting, less monotonous and, above all, unpredictable to the listener. Finally, six other items were selected for a training session; their structure covered all the types used (target, empty, early, late). In sum, 118 stimuli were presented to the participants, who needed from 15 to 19 minutes to complete the session.

4.1.3 Listeners and procedure

The experiment was administered to 50 native speakers of Czech (18–53 years, mean = 24 years). There were 30 women and 20 men, and all but three came from Prague or Central Bohemia, most of them students of various universities.

The sessions took place in 2021 at the Institute of Phonetics, Charles University, Prague, in a sound-treated room without any disturbances or interruptions. The experiment was run in *Dmdx* (Forster & Forster, 2003; see Šturm & Volín, 2012). The sound level was first adjusted according to the participant's needs. The experiment started with written instructions specifying the task and procedure. Each item consisted of the following: (1) the prompt, displayed in the centre of the screen in uppercase letters; (2) a short audio signal (beep) followed by silence, indicating the start of the stimulus; (3) the stimulus itself. The participants were instructed to press a large button in front of them as soon as they heard and recognized the word presented orthographically (see Kilborn & Moss, 1996, for a methodological discussion of the task). The button (Black Box Toolkit) was designed specifically for RT measurements with millisecond accuracy. When the sentence did not include the target word, the participant just waited. No other action was needed, and the next item appeared automatically.

After a training session, the 112 remaining stimuli were presented in four blocks (with a short, one-minute break between them). Each block contained 20 target items and 8 fillers. The order of items within a block was randomized for each participant. However, the order of blocks was semi-random due to the fact that pairs of stimuli (glottalization vs. linking) were used. Therefore, a glottalized version of a carrier could not appear in the same half as the counterpart version (in other words, the same text was restricted to different halves). To prevent any order effects, two versions of the experiment were created, balanced across participants. In one, blocks A and B (in random order) were followed by blocks C and D (in random order). In the other version, blocks C and D preceded blocks A and B.

4.1.4 Data processing and analysis

There were 4000 target observations (without fillers). As is common practice in RT research, two types of reactions were dismissed: anticipations (button is pressed sooner than the brain's cognition allows for a proper response) and misses/delays (button is pressed later than what is regarded as immediate reaction). It is clear that defining a valid response is not a simple task, and the decision differs across studies (see Jiang, 2012). In line with Bissiri et al. (2011), we defined a 'hit' as a response between 150 ms and 1000 ms. As a result, 3617 responses remained in the dataset. Subsequently, unreliable listeners and items were discarded (for simple tasks, Jiang, 2012 recommends dismissing listeners with an error rate over 20%). Two listeners were discarded (error rate 23%), and three pairs of items (error rate 80–90%; the remaining items had error rates below 13%). The final dataset comprised 3466 observations.

As is typical in RT research, histograms revealed a highly positive skew in the RT distribution. Therefore, the RT values in milliseconds were logarithmically transformed, which resulted in a normal distribution.

Statistics and data visualization was performed in *R* (R Core Team, 2020), using the libraries *lme4* (Bates, Mächler, Bolker & Walker, 2015), *emmeans* (Lenth, 2022) and *ggplot2* (Wickham, 2009). LME models were created separately for the two types of manipulation, with glottalization (linked × glottalized) and distance (1, 2, 3, 4, 5-7 syllables) as fixed effects and listener and prompt as random effects. The distance was a factor variable, not numeric.

4.2 Results

A simple comparison of median values for glottalized and non-glottalized conditions (Fig. 2) revealed a negligible difference of 0.008 on the logarithmic scale (translating to approx. 3 ms after back transformation, or less than 1%). Since the two types of manipulation could yield different results, Figure 3 splits the dataset according to whether glottalization was added or eliminated. However, very similar values appeared in each condition (a change of 3 ms, or 1%, for addition, a change of 10 ms, or 3%, for elimination). The violin plots also confirm the normal distribution of log-transformed values.

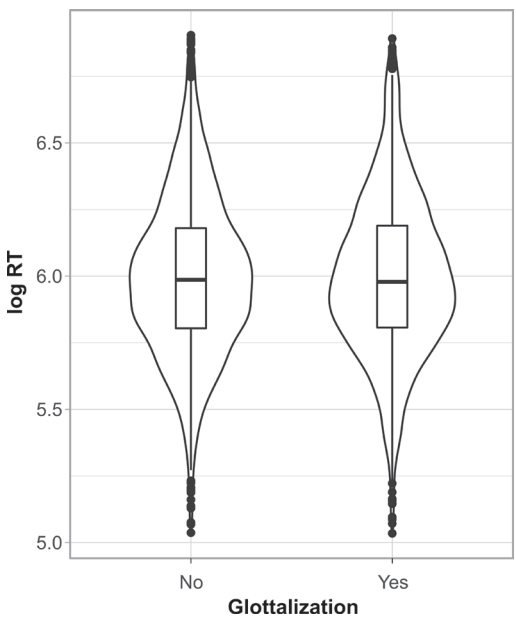

Figure 2. Logarithmic RTs of glottalized and linked stimuli in the full dataset.

Although glottalization does not seem to elicit any substantial RT difference in the participants' behaviour, we must still consider the effect of the distance between the prompt and the carrier. Moreover, the graphs so far described independent data. Figure 4 now displays paired differences between the two versions of each stimulus, where positive values mean shorter RTs in the glottalized version. Only boxplots are shown for better

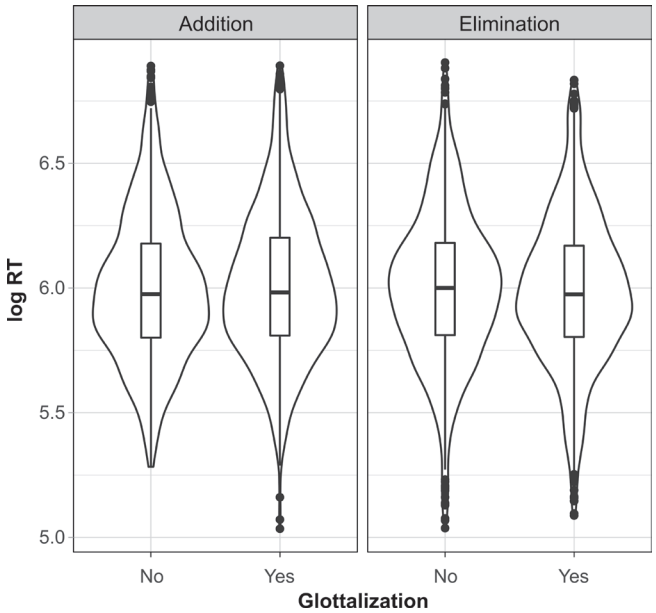

Figure 3. Logarithmic RTs of glottalized and linked stimuli in two types of manipulation. Note that "No" under Addition and "Yes" under Elimination mark the original as compared to the manipulated stimulus.

clarity. The graph on the left suggests that addition had little impact on the RTs, and with distances of 3 and 4 syllables even seemed to produce the opposite effect (higher RTs of glottalized stimuli). In contrast, elimination yielded either small effects or, with distances of 1 and 3 syllables, an effect in the predicted direction (a change of 28 ms, or 7%, for the immediate distance; 15 ms, or 4%, for the distance of 3 syllables). The results will be evaluated below in a statistical model.

Two LME models were created. First, using the addition of glottalization dataset, the manipulation condition and the distance of prompt from target were not significant ($\chi^2(1) = 2.1$, $p = 0.147$; $\chi^2(4) = 5.3$, $p = 0.253$). Likewise, there was no significant interaction between the two effects ($\chi^2(4) = 5.0$, $p = 0.287$). However, in the second model (the elimination dataset), there was a significant interaction of manipulation * distance ($\chi^2(4) = 21.9$, $p < 0.001$). Specifically, post-hoc tests revealed a significant effect in the immediate condition one syllable from the target (Linked – Glottalized = 0.1024, SE = 0.0237, t-ratio = 4.327, $p < 0.001$). After back transformation, such an effect corresponds to 40 ms (or a change of 11%). No other distances were associated with significantly different pairs ($p > 0.05$). The effect plots are shown in Figure 5.

4.3 Discussion

The aim of our study was to contribute to the research into the functionally-defined phenomenon of boundary glottalization, mostly in relation to its impact on the cognitive load of native Czech listeners. This is an extension to previous research (Bissiri et

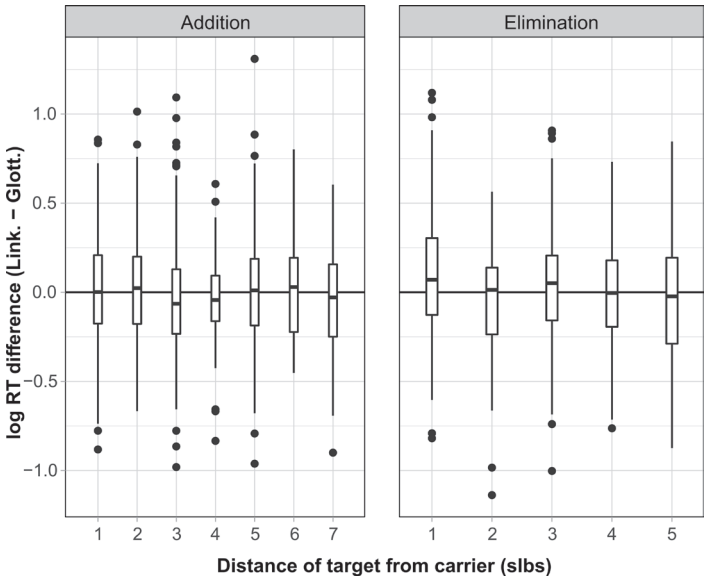

Figure 4. A difference in logarithmic RTs of glottalized and linked stimuli in two types of manipulation as a function of distance. Positive values indicate shorter RTs in the glottalized stimuli.

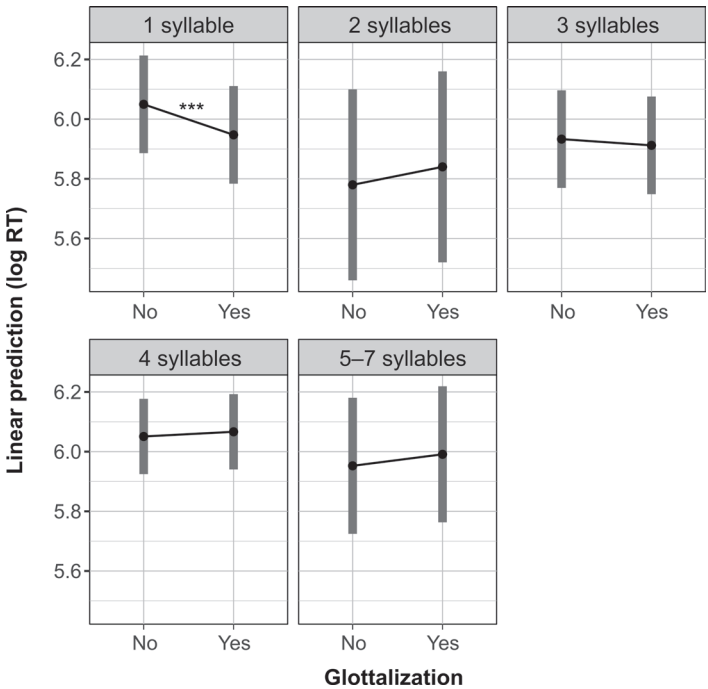

Figure 5. Effect plots of the second LME model (elimination of glottalization manipulation) with an interaction between glottalization and distance of prompt word from carrier word.

al., 2011; Volín et al., 2012; Schwartz et al., 2015), which used non-native listeners as participants. The hypothesis was that the presence of glottalization in the stimulus lowers the cognitive load, and thus speeds up the processing of the following words. To test this prediction, we measured reaction times (RTs) in a word-monitoring experiment. In contrast to the previous studies, we used a more spontaneous material (TV debate) and we also focused on the distance of the glottalization from the target word, which makes our research innovative.

The results of the behavioural experiment partly confirmed the prediction. Stimuli with glottalization were associated with shorter RTs than stimuli with linking, but only when the target word (the prompt with an initial consonant) appeared one syllable after the manipulated carrier context. The cognitive benefit of glottalization thus seems to be restricted to immediate surroundings, and is not extended to more distant words. This would be in line with the results of the previous studies, which used an even shorter distance (zero syllables: the prompt was identical to the carrier word). Moreover, the effect was apparent only in one of two types of manipulation (elimination of glottalization rather than its addition).

Interestingly, upon a reviewer's question regarding gender differences, we examined post hoc whether male and female participants behaved similarly with respect to the glottalization effect. Although male participants were somewhat faster, which is in accordance with previous research (e.g., Jain, Bansal, Kumar & Singh, 2015), we did not find any reliable differences in the performance of the two groups in relation to the presence of glottalization. Therefore, glottalization seems to be perceived similarly by both genders, in contrast to its production (usage).

However, the merit of our research cannot be reduced to the results only; what seems to be equally valuable are methodological observations from the design of the experiment. Although the use of material of a lower level of control has its justification in the attempt to approximate common communicative situations, our experiment has shown that employing such a type of material in a perception test offers several obstacles which may, without sufficient attention, reduce the validity of the experiment. One of the pitfalls is the stimuli selection process, which is restricted by the available recordings and the factors included in them. More attention and control should be paid to the key factors, such as balancing the number of stimuli with respect to prompt distance, stimulus length, or the semantic status of the carrier words. Such a process will place higher demands on the extent of available material and the researcher's time.

Another suggestion for future research relates to the creation of the manipulated stimuli. Such a process may disturb the temporal characteristics of the original recordings. It is a well-known fact that an interference with the rhythm of speech increases by itself the cognitive load of processing. Despite our attempts to minimize the temporal changes introduced by manipulation, it is important to bear in mind that manipulated and original items automatically have a different starting point, regardless of the level of the independent variable (linked vs. glottalized in our case). Results could then be biased or misinterpreted. In fact, the hypothesis that glottalization leads to shorter RTs should be supplemented with the hypothesis that manipulation leads to longer RTs. In this light, our results (and those of Bissiri et al., 2011 and Volín et al., 2012) make more sense. The addition of glottalization combines two opposing effects, which produced null or very

weak outcomes, whereas the elimination of glottalization combines two parallel effects, which yielded outcomes of different strength in the expected direction. Solution to this methodological problem is an important task for researchers in the field of RT measurement. An obvious escape route – presenting natural, non-manipulated stimuli, as in Schwartz et al. (2015) – does not seem to be the way to go. Their results showed only weak effects, which might be due to the fact that the paired versions were not completely identical. Only a single point of difference (glottalization) is clearly needed between the matched versions.

5. Conclusions

According to our hypothesis, the presence of word-initial glottalization should decrease the cognitive difficulty of sentence processing, leading to a decrease in reaction times (RTs) for glottalized items. This prediction was confirmed partially, as the effect was restricted to immediate contexts (distance of one syllable between the prompt word and the carrier word) and to one type of manipulation (elimination as opposed to addition of glottalization). The experiment with native listeners thus confirmed some of the previous results from L2 processing, and highlighted certain methodological aspects of an RT experiment design. Finally, the experiment showed both merits and difficulties of using speech material characterized by a lower level of control than the commonly used read speech.

Acknowledgements

This study was supported from the European Regional Development Fund-Project 'Creativity and Adaptability as Conditions of the Success of Europe in an Interrelated World' (No. CZ.02.1.01/0.0/0.0/16_019/0000734) and was created within the programme 'Cooperatio', scientific field Linguistics.

REFERENCES

Ashby, M. & Przedlacka, J. (2011). The stops that aren't. In: *English Phonetics 14–15. Festschrift commemorating the retirement and the 70th birthday of Professor John C. Wells*. The English Phonetic Society of Japan.

Ashby, M. & Przedlacka, J. (2014). Measuring incompleteness: Acoustic correlates of glottal articulations. *Journal of the International Phonetic Association*, 44(3), 283–296.

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67, 1–48.

Bissiri, M. P., Lecumberri, M. L., Cooke, M. & Volín, J. (2011). The role of word-initial glottal stops in recognizing English words. In: *Proceedings of Interspeech 2011*, Florence, pp. 165–168.

Boersma, P. & Weenink, D. (2021). *Praat: Doing phonetics by computer (version 6.1)* [Computer software]. Retrieved from <http://www.praat.org/>.

Cruttenden, A. (2014). *Gimson's Pronunciation of English* (8th ed.). Taylor and Francis.

Di Napoli, J. (2015). Glottalization at phrase boundaries in Tuscan and Roman Italian. In: J. Romero & M. Riera (Eds.), *The Phonetics–Phonology Interface: Representations and Methodologies* (pp. 125–147). John Benjamins.
Dilley, L., Shattuck-Hufnagel, S. & Ostendorf, M. (1996). Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetics*, 24, 423–444.
Docherty, G. & Foulkes, P. (1999). Sociophonetic variation in 'glottals' in Newcastle English. In: *Proceedings of the 14th International Congress of Phonetic Sciences*, San Francisco, pp. 1037–1040.
Fabricius, A. (2002). Ongoing change in modern RP: Evidence for the disappearing stigma of t-glottalling. *English World-Wide*, 23(1), 115–136.
Forster, K. I. & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods, Instruments & Computers*, 35(1), 116–124.
Gavaldà, N. (2016). Individual variation in allophonic processes of /t/ in Standard Southern British English. *The International Journal of Speech, Language and the Law*, 23(1), 43–69.
Gobl, C. & Ní Chasaide, A. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*, 40(1–2), 189–212.
Gordon, M. & Ladefoged, P. (2001). Phonation types: A cross-linguistic overview. *Journal of Phonetics*, 29, 383–406.
Grászi, T. E. & Markó, A. (2018). Word-initial glottal marking in Hungarian as a function of articulation rate and word class. In: *Challenges in Analysis and Processing of Spontaneous Speech* (pp. 75–98). mta Nyelvtudományi Intézet.
Grossberg, S. (2003). Resonant neural dynamics of speech perception. *Journal of Phonetics*, 31(3–4), 423–445.
Hála, B. (1967). *Výslovnost češtiny I*. Academia.
Jain, A., Bansal, R., Kumar, A. & Singh, K. D. (2015). A comparative study of visual and auditory reaction times on the basis of gender and physical activity levels of medical first year students. *International Journal of Applied and Basic Medical Research*, 5(2), 124–127.
Jiang, N. (2012). *Conducting Reaction Time Research in Second Language Studies*. Routledge.
Keating P., Garellek, M. & Kreiman, J. (2015). Acoustic properties of different kinds of creaky voice. In: *Proceedings of the 18th International Congress of Phonetic Sciences*, Glasgow, paper 821.
Kilborn, K. & Moss, H. (1996). Word monitoring. *Language and Cognitive Processes*, 11(6), 689–694.
Kohler, K. (1994). Glottal stops and glottalization in German. *Phonetica*, 51, 38–51.
Kopečková, M. (2020). *Analýza zvukové roviny mluvního projevu moderátorů hlavního TV zpravodajství* [Ph.D. thesis]. Univerzita Palackého v Olomouci, Filozofická fakulta.
Ladefoged, P. & Maddieson, I. (1996). *The Sounds of the World's Languages*. Blackwell.
Lenth, R. V. (2022) *emmeans: Estimated Marginal Means, aka Least-Squares Means* (1.7.2) [Computer software]. Retrieved from <https://CRAN.R-project.org/package=emmeans>.
Local, J., Wells, W. H. G. & Sebba, M. (1985). Phonology for conversation: Phonetic aspects of turn delimitation in London Jamaican. *Journal of Pragmatics*, 9, 309–330.
Maddieson, I. (1984). *Patterns of Sounds*. Cambridge University Press.
Malisz, Z., Żygis, M. & Pompino-Marschall, B. (2013). Rhythmic structure effects on glottalisation: A study of different speech styles in Polish and German. *Laboratory Phonology*, 4(1), 119–158.
Ogden, R. (2001). Turn transition, creak and glottal stop in Finnish talk-in-interaction. *Journal of the International Phonetic Association*, 31(1), 139–152.
Olaya, N. P. (1967). *A Phonological Grammar of a Dialect of Ilokano* [M.A. thesis]. University of British Columbia.
Palková, Z. (2016). Hlasivkový ráz ve výzkumu zvukové stavby češtiny. In: A. Cychnerska & I. Sawicka, *Sandhi w językach słowiańskich II* (pp. 143–158). Wydawnictwo Naukowe Uniwersytetu Mikołaja Kopernika.
Pompino-Marschall, B. & Żygis, M. (2010). Glottal marking of vowel-initial words in German. *ZAS Papers in Linguistics*, 52, 1–17.
R Core Team (2020). *R: A language and environment for statistical computing (Version 4.0.3)* [Computer software]. R Foundation for Statistical Computing. Retrieved from <https://www.r-project.org>.
Redi, L. & Shattuck-Hufnagel, S. (2001). Variation in the realization of glottalization in normal speakers. *Journal of Phonetics*, 29, 407–429.

Schwartz, G., Rojczyk, A. & Balas, A. (2015). Monitoring English sandhi-linking: A study of Polish listeners' L2 perception. *Research in Language*, 13(1), 61–76.
Skákal, L. (2013). *Užívání hlasivkového rázu u rodilých a nerodilých mluvčích francouzštiny* [B.A. thesis]. Univerzita Karlova, Filozofická fakulta.
Skákal, L. (2015). *Užívání glotalizace jako faktor umožňující identifikaci mluvčího* [M.A. thesis]. Univerzita Karlova, Filozofická fakulta.
Skarnitzl, R. (2004). Acoustic categories of nonmodal phonation in the context of the Czech conjunction "a". In: Z. Palková & J. Veroňková (Eds.), *AUC Philologica 1/2004, Phonetica Pragensia X* (pp. 57–68). Karolinum.
Skarnitzl, R., Čermák, P., Šturm, P., Obstová, Z. & Hricsina, J. (2021). Glottalization and linking in the L2 speech of Czech learners of Spanish, Italian and Portuguese. *Second Language Research*, 38(4), 941–963. DOI: 10.1177/02676583211015803.
Šturm, P. & Bičan, A. (2021). *Slabika a její hranice v češtině*. Karolinum.
Šturm, P. & Volín, J. (2012). Měření reakčních dob u experimentů s akustickými podněty. *Akustické listy*, 18(2–4), 25–30.
Veroňková J. (2016). Výskyt rázu ve čtených a semispontánních projevech v češtině. In: A. Cychnerska & I. Sawicka, *Sandhi w językach słowiańskich II* (pp. 159–172). Wydawnictwo Naukowe Uniwersytetu Mikołaja Kopernika.
Volín, J. (2012). Jak se v Čechách „rázuje". *Naše řeč*, 95(1), 51–54.
Volín, J., Uhrinová, M. & Skarnitzl, R. (2012). The effect of word-initial glottalization on word monitoring in Slovak speakers of English. *Research in Language*, 12(2), 173–181.
Wagner, P., Trouvain, J. & Zimmerer, F. (2015). In defense of stylistic diversity in speech research. *Journal of Phonetics*, 48, 1–12.
Wickham H. (2009). *Ggplot2: Elegant Graphics for Data Analysis*. Springer.
Wolk, L., Abdelli-Beruh, N. B. & Slavin, D. (2012). Habitual use of vocal fry in young adult female speakers. *Journal of Voice*, 26(3), e111–e116.
Yuasa, I. P. (2010). Creaky voice: A new feminine voice quality for young urban-oriented upwardly mobile American women? *American Speech*, 85(3), 315–337.

Appendix

Target items used in the perception test. All caps indicate the prompt (target word) displayed on the screen. Carrier words in bold are sites of glottalization where manipulations were performed.

Originally glottalized items:	Originally linked items:
1) *by si měl, nebo si **uvědomil** ZŘEJMĚ, a přišel taky*	21) *taková by měla být **omezena** POUZE na projevy učiněné v souvislosti*
2) *je, tak by měl kopírovat **agendu** MINISTERSTVA kultury*	22) *poslanci by neměli mít **imunitu**, DISKUTOVAL bych o tom*
3) *to, že ti **občané** MUSÍ mít pocit*	23) *má být **ošetřováno** RŮZNĚ jízdné a podobně*
4) *najevo. Pan Milota skončil **okamžitě** V PODSTATĚ ve funkci*	24) *priority jsou tak **odlišné**, že KAŽDÝ ví, že*
5) *si **uvědomil**, že jeho MÍSTO je jinde a setsakramentsky*	25) *základní pilíře, prevence, průhlednost **a** POSTIH. Všechny veřejné*
6) *nikoliv **už** za MÍSTOPŘEDSEDU poslaneckého výboru*	26) *zástupci z řad **občanů**, to ZNAMENÁ bude vytvořen registr*
7) *Jsem přesvědčen **o** tom že PROSTĚ*	27) *se může přihlásit **a** BUDE přímo u zadávání*

Originally glottalized items:	Originally linked items:
8) *vás těším **opět** PŘÍŠTÍ neděli v jedenáct hodin*	28) *Bude moct **účinně** PROTI korupci bojovat*
9) *tu vládu **asi** víc TLAČIT, protože již předminule*	29) *pochybnost, že prostě nějaký **úředník** ve SKLEPENÍ magistrátu dělá něco*
10) *chystali do **opozice** po KVĚTNOVÝCH volbách sněmovních*	30) *pomalu, ta **otázka** ZNĚLA, jestli bude*
11) *my nebudeme nikoho kádrovat, **abychom** ŘÍKALI, že ten je vhodný*	31) *v další části budeme mluvit **o** BOJI proti korupci*
12) *nevšiml jsem si, že by **akciová** společnost FUTURA měla nějak*	32) *prezidentský kandidát **udělá** KAMPAŇ, jakou uzná*
13) *ten problém, **o** KTERÉM se tolik nemluví*	33) *v pátek jednání s **odboráři** DOPRAVNÍHO podniku*
14) *to měl být, mně přijde **úplně** LICHÉ, protože*	34) *několika dnů vyloučeni z **občanské** DEMOKRATICKÉ strany*
15) *každý **občan** ČESKÉ republiky má být právo*	35) *se to takhle říct **od** STOLU, protože já*
16) *to, **abychom** se jako POLITICKÉ strany dohodli*	36) *určitě ano, jak **imunita**, tak POSLANECKÉ náhrady*
17) *bych měl jít příkladem, **a** NEMŮŽEME od lidí něco chtít*	37) *je to například na **úřadech**, že JEDNOU za rok*
18) *ti kteří vlastně o těch zakázkách **a** MILIARDÁCH a směřování*	38) *něco jako **omezení** HORNÍ hranice*
19) *dnes pozice **imunity** ZTRATILA svůj význam*	39) *Že zástupce komunistické strany se **odvolává** na RYCHLÉ šípy a na na skauty*
20) *Dalších **institucí** je NÍZKÁ i z těchto důvodů*	40) *bych se ani nebavil **o** těch NÁKLADECH na kampaň, protože*

Michaela Svoboda
Institute of Phonetics
Faculty of Arts, Charles University
Prague, Czech Republic
E-mail: michaela.svoboda@post.cz

Pavel Šturm
Institute of Phonetics
Faculty of Arts, Charles University
Prague, Czech Republic
E-mail: pavel.sturm@ff.cuni.cz

2022 ACTA UNIVERSITATIS CAROLINAE PAG. 51–63
PHILOLOGICA 1 / PHONETICA PRAGENSIA

THE DYNAMIC EFFECT OF SPEAKING FAST ON SPEECH PROSODY

LAURI TAVI

ABSTRACT

Speaking fast causes several changes in speech prosody. In addition, it can be associated with a decrease in speech intelligibility. In this study, prosodic changes in fast speech were investigated using common prosodic measurements and syllabic prosody index (SPI), a novel prominence measure that combines f0, energy and duration features. Dynamic changes in long-term prosodic prominence were investigated using functional data analysis (FDA), in which the SPI is transformed into a functional form. The possibly decreasing effect of speaking fast on speech intelligibility was evaluated using automatic speech recognition. Phonetic analyses of syllabic units showed that speaking fast decreases duration, f0 and SPI, and increases articulation rate and proportional acoustic energy in the frequency range of 0–1 kHz. FDA supported the aforementioned results by revealing dynamically decreased overall prominence in fast speech. Furthermore, in comparison to regular speech, speech intelligibility was found to be significantly lower in fast speech: word error rate (WER) for regular speech was 0.27, whereas for fast speech it was 0.86.

Keywords: fast speech, prosody, prominence, functional data analysis, speech intelligibility

1. Introduction

It is well known that speech characteristics, such as prosodic features and articulatory gestures, change dynamically over time (Roettger et al., 2019; Niebuhr et al., 2011). The dynamic changes occur particularly in natural speech communication, in which speakers alter their speech both voluntarily and involuntarily. According to Lindblom's theory of Hyper and Hypoarticulation (H&H), for example, speakers intentionally adapt their speech according to conversational demands (Lindblom, 1990). In other words, speakers' articulatory effort can decrease (hypoarticulation) or increase (hyperarticulation) depending on how intelligible they believe their speech is for listeners. An example of an involuntary change in prosody is the Lombard effect, which causes speakers to increase their vocal effort in noisy conditions (Stanton et al., 1988; Patel & Schell, 2008). These changes include increases of pitch and duration of words, yielding improved speech audibility and intelligibility.

Some prosodic changes, such as decrease in word duration, can also decrease speech intelligibility (Mayo et al., 2012; Hazan & Markham, 2004). In addition to the

https://doi.org/10.14712/24646830.2022.28

fact that words may be less carefully articulated in fast speech (Janse, 2004), the timing patterns are different from those in regular speech tempo. When speaking fast, the duration of unstressed syllables is reduced more than that of stressed syllables, resulting in a more prominent prosodic pattern (Janse, 2004). The more prominent pattern in fast speech, however, probably does not improve intelligibility. In fact, increasing the speech rate artificially without changing prominence patterns has yielded speech that is more difficult to process compared to naturally fast speech (Janse et al., 2003; Janse, 2004).

Prosodic prominence refers to the relative emphasis of syllables, which can be acoustically measured as the variation of relative energy, duration and fundamental frequency (f0) (Greenberg et al., 2003; Tavi & Werner, 2020). Although the term "prominence" typically refers to relative changes, for example, between adjacent syllables, here the term is also used to describe the strength of emphasis between different speaking conditions. During fast speech, speakers might not be able to properly emphasize syllables due to the limited articulation and processing time, which could result in a decrease of overall prominence. However, to the author's knowledge, previous phonetic studies lack acoustically orientated analyses of exactly how fast speech impacts the dynamics of prosodic prominence in healthy speakers.

In order to establish the relationship between fast speech and prosodic prominence, two hypotheses were formulated and tested in this study: Speaking fast (1) decreases prosodic prominence and (2) impairs speech intelligibility. The possible decrease of prosodic prominence is examined focusing on different aspects of prosody, i.e., pitch, energy and durational characteristics. To avoid subjective listening tests, speech intelligibility was evaluated using the accuracy of automatic speech recognition (ASR) in terms of word error rate (WER). WER has been a common metric to evaluate speech intelligibility in an objective and comparable manner in numerous technology-oriented speech studies, such as in Voice Privacy Challenge (Tomashenko et al., 2021).

In addition to inspecting conventional statistics, this study utilizes functional data analysis (FDA) in order to reveal wide-scale dynamic differences in prosodic prominence between regular and fast speech. The focus is on the steadiness, or major shifts, of long-term prominence rather than on high-frequency prominence variation between adjacent syllables. FDA is a methodology which extends conventional statistics from discrete values to functions of time (Ramsay et al., 2009). One popular method in FDA has been functional principal component analysis (fPCA), in which eigenvalues are paired with eigenfunctions instead of eigenvectors as in traditional PCA. In previous phonetic studies, fPCA has been shown to be an effective method of capturing the dynamic nature of speech (Cronenberg et al., 2020; Gubian et al., 2010, 2011; Zellers et al., 2010; Gubian et al., 2015).

In this paper, Section 2 will describe the speech data and analysis techniques used in this study. The answers to the aforementioned hypotheses will be presented in Section 3 and discussed in Section 4.

2. Speech data and methods

2.1 The Chains corpus

Prosodic characteristics of fast speech are investigated using the Chains corpus. The Chains corpus was collected in 2005 in order to study challenges in speaker identification (Cummins et al., 2006). The corpus contains six different speaking conditions (i.e., retelling, synchronous imitation, repetitive synchronous, solo, fast, and whispered speech) from 36 (20 male and 16 female) speakers. The speakers read aloud four short fables (”Cinderella”, ”Rainbow text”, ”North Wind and the Sun”, and ”Members of the Body”) and 33 individual sentences. In this study, only the four fables produced with solo (hereafter referred to as “regular”) and fast speaking conditions were analysed.

2.2 Phonetic measurements

Phonetic analyses were carried out with Praat (Boersma and Weenink, 2020) and performed separately for the female and male speakers. Firstly, the readings of the four fables were automatically segmented into syllabic units using Vocal toolkit (Corretge, 2020), which adapts a script (De Jong and Wempe, 2009) to detect syllable nuclei. The term ”syllabic unit” is used here because automatic syllable markings are not perfectly aligned with linguistic syllables.

Secondly, a total of five acoustic-phonetic features were analysed from the syllabic units: articulation rate, duration, relative energy proportion below 1 kHz in a frequency range of 0–4 kHz, median f0 and syllabic prosody index (SPI). The SPI (Tavi and Werner, 2020) measures prosodic prominence in syllables by combining their pitch, duration and energy proportion below 1 kHz into one feature. SPI is formulated as

$$SPI = \frac{Pitch_{median} \times \sqrt{Duration}}{\sqrt{Energy_{below1kHz}}}/10.$$

The higher the SPI, the higher the prominence in a syllabic unit. In pitch analysis, the ceiling and the floor values were set to 120 and to 500 Hz for female speakers and to 70 and to 400 Hz for male speakers. The relative energy proportion was calculated by dividing the overall energy in the frequency range of 0–1 kHz by the overall energy in the frequency range of 0–4 kHz. This measure of spectral tilt is considered as an indicator of emphasis, since in weaker speech segments, energy is more concentrated in the lower frequencies (Tavi & Werner, 2020).

2.3 Functional data analyses

In the first step of FDA, scalar SPI values were transformed into logarithmic continuous functions, or functional SPIs (fSPIs). A (natural) logarithmic scale was used due to the fact that speech perception is also logarithmic (Reetz, 2009). Only the SPI measurements were used in functional analyses because they present all the main prosodic features (Tavi & Werner, 2020) in a single measure. The B-spline basis system was used for

transforming the scalar SPI values to fSPIs, as it is a common choice for aperiodic signals (Gubian et al., 2015). The order of polynomial segments was set to four and the number of basis functions was set to 42. The lambda parameter, which defines the amount of smoothness, was 0.1. These parameters were chosen based on visual inspections of the resulting functions using different levels of smoothing. A strong smoothing was applied in order to take into account only the major variation in prosodic prominence and to exclude short-term fluctuation in adjacent syllabic units.

Because the aim of this study was to analyse prosodic prominence in different speaking conditions rather than specific linguistic phrases, the mean fSPI of the four fables was calculated for each speaker for both regular and fast speech. As a result, the mean fSPIs carry information on averaged wide-scale prosodic events, in which the amount of intra-speaker and linguistic variation has been reduced.

Finally, fPCA was applied on the mean fSPIs (hereafter referred to as fSPIs instead of mean fSPIs). Using fPCA, fSPIs can be reconstructed with the formula:

$$f(t) \approx \mu(t) + \sum_{i=1}^{n} s_i \times PC_i(t),$$

where $\mu(t)$ is the mean of the fSPIs, PC_i is the ith principal component function, and s_i is its weight, or score. Because the individual scores model the shape of each function, they can be used to investigate the dynamics of continuous speech features, such as f0 or formant curves (Gubian et al., 2015).

3. Results

3.1 Prosodic features

A total of five prosodic features were extracted from the syllabic units. The measurements were compared using paired T-tests. Articulation rate was measured in order to confirm that syllabic units are truly spoken faster in fast compared to regular speech.

Table 1. Prosodic differences between regular and fast speaking conditions presented as mean values and p-values from paired T-tests (For 'feature' see Section 2.2).

feature	*female speech*					*male speech*				
	regular	*fast*	*t*	*df*	*p*	*regular*	*fast*	*t*	*df*	*p*
AR	5.06	5.86	11.35	15	<.001	4.95	5.95	–14.36	19	<.001
f0 (Hz)	199	190	4.27	15	<.001	120	110	6.17	19	<.001
Eb1kHz	0.91	0.99	–8.74	15	<.001	0.93	0.99	–10.86	19	<.001
dur (s)	0.20	0.17	11.46	15	<.001	0.20	0.17	15.63	19	<.001
SPI	9.19	7.77	8.50	15	<.001	5.51	4.45	13.30	19	<.001

Table 1 shows the mean values of the two speaking conditions calculated from all speakers and the results from the paired T-tests. The results are presented separately for male and female speakers. In comparison to regular speaking condition, speaking fast increased the articulation rate and energy proportion below 1 kHz. Duration, f0 and SPI in syllabic units were decreased. The prosodic differences between the speaking conditions were statistically significant for both the male and the female speakers using a Benjamini & Hochberg -adjusted significance level of 0.05. The results confirmed that syllabic units were spoken faster in fast speech and that increasing speech tempo has a significantly decreasing effect on prosodic prominence.

3.2 Functional syllabic prosody index

In order to examine the dynamic changes in prosodic prominence caused by fast speaking, the SPI trajectories were converted to functional SPIs (see Section 2.3). Figure 1 shows the mean fSPIs for male and female speakers. The clearest difference between the regular and the fast mean fSPIs for both male and female speakers is that the regular fSPIs are above the fast fSPIs. The lower fSPIs for fast speech were expected based on the results of the acoustic-phonetic analyses presented in Section 3.1. The positions of the mean fSPIs are rather consistent throughout the whole functions, indicating that speakers are able to retain constant prominence levels in long speech segments of different tempi.

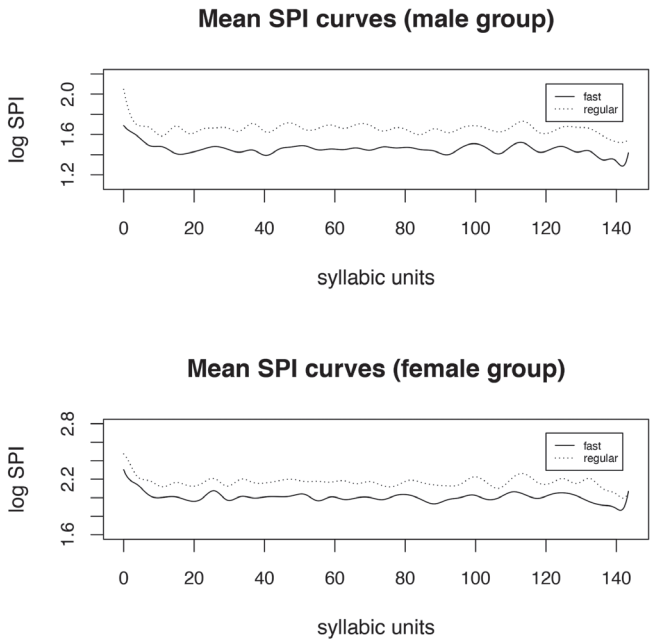

Figure 1. Mean fSPIs for male and female speakers.

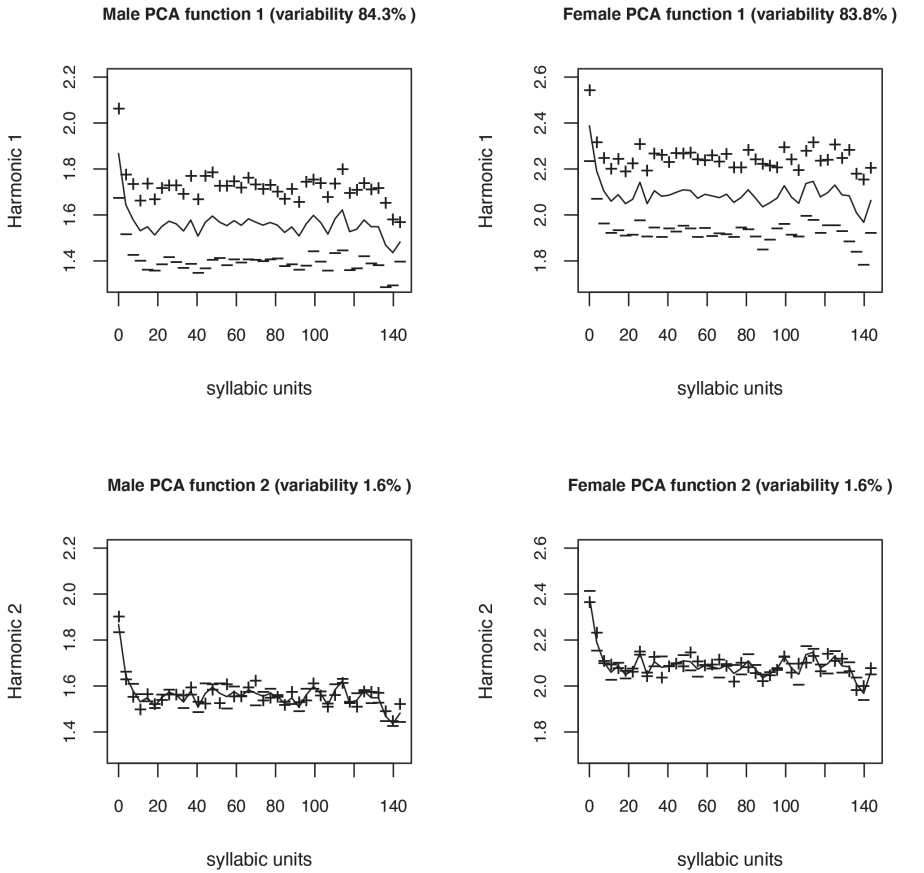

Figure 2. The effect of the first two PC functions on the male and the female speakers' mean fSPIs. Trajectories of plus and minus signs demonstrate the effects of the PC functions and standard deviation of their scores on the mean fSPIs (solid lines).

fPCA was applied to the fSPIs in order to examine the major modes of prominence variation. Figure 2 demonstrates the effects of PC functions on the mean fSPIs using the trajectories of plus and minus signs. These trajectories are formed by multiplying the PC functions by the standard deviation of their weightings, which are then either added to or subtracted from the mean fSPI. The effects are rather similar for male and female speakers: PC1 (top panels) explains the variation related to location of the fSPI. It also explains a major part of the fSPI variability (>80%). An increase of PC1 weighting, or s_1, will raise the mean fSPI, whereas decrease of the weighting will lower it. PC2 and its weighting, or s_2, are more associated with the timings of positive and negative prominence peaks; however, the second PCA function explains only 1.6% of the fSPI variation.

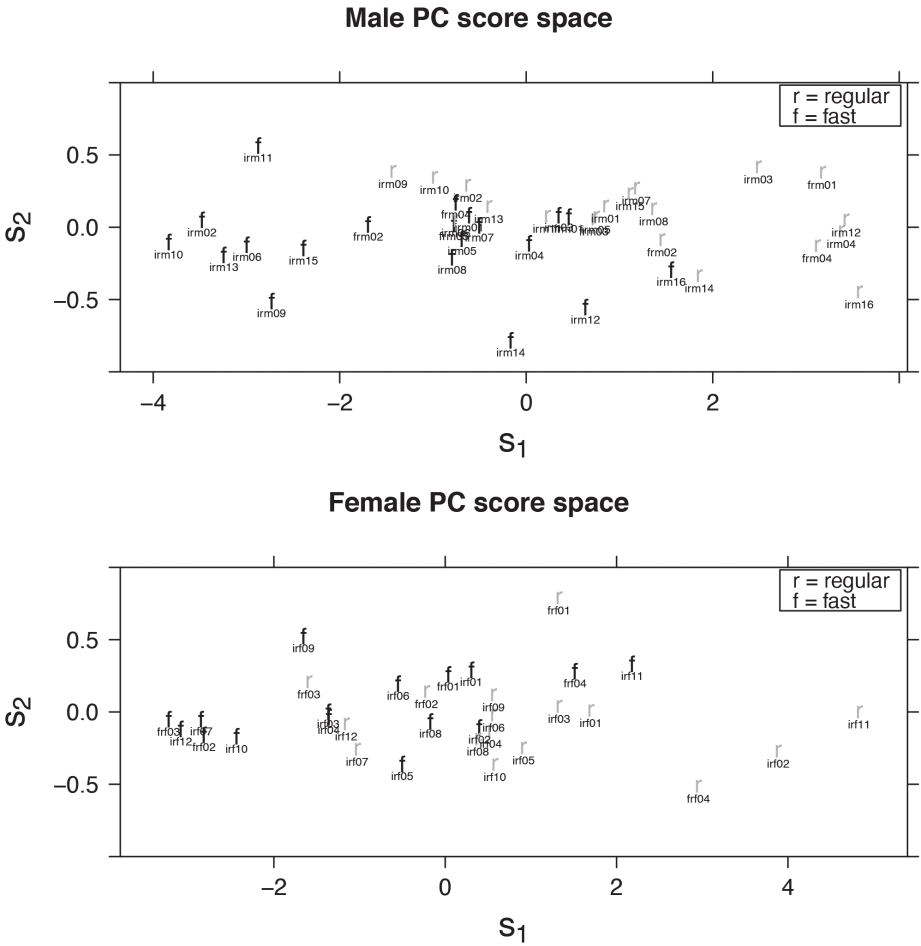

Figure 3. PC score spaces for male and female speakers.

Figure 3 reveals how the two speaking conditions of individual speakers are located in the PC1–PC2 score space. Letters *f* (fast) and *r* (regular) indicate the speaking condition and the identifier below the letters indicates speaker identity. In the PC score spaces, fast speech is mainly located on the left and regular speech on the right. The division between the speaking conditions is clearer with the male speakers, as the female speakers' scores have more overlap. However, each speaker's s_1 of fast speech is lower compared to their s_1 of regular speech (see Figure 4). The s_2 shows no relationship specific to the two speaking conditions.

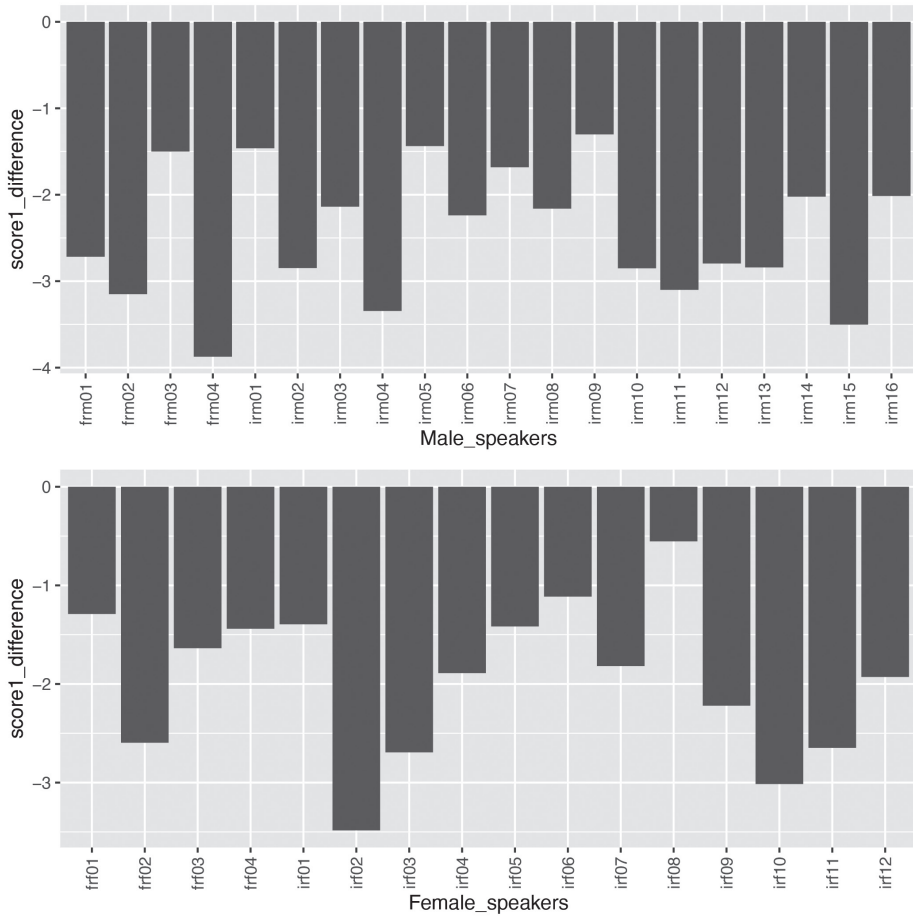

Figure 4. The s_1 differences between fast and regular speech for each speaker. The s_1 of regular speech is subtracted from the s_1 of fast speech.

3.3 Speech intelligibility

Sections 3.1 and 3.2 described several changes in prosody caused by the increase of speech tempo. To test whether these changes are related to speech intelligibility, the four fable passages were transcribed using an ASR system provided by BAS Web Services (Kisler et al., 2017). Then, a Python script[1] was used to calculate WERs for each speaker's regular and fast versions of the passages. WER divides the number of errors (i.e., the substitutions, insertions and deletions) by the total number of words. Although WER is reported as a percentage, it can be more than 100%, because the number of errors can be higher than the number of words in the reference text.

[1] https://holianh.github.io/portfolio/Cach-tinh-WER/

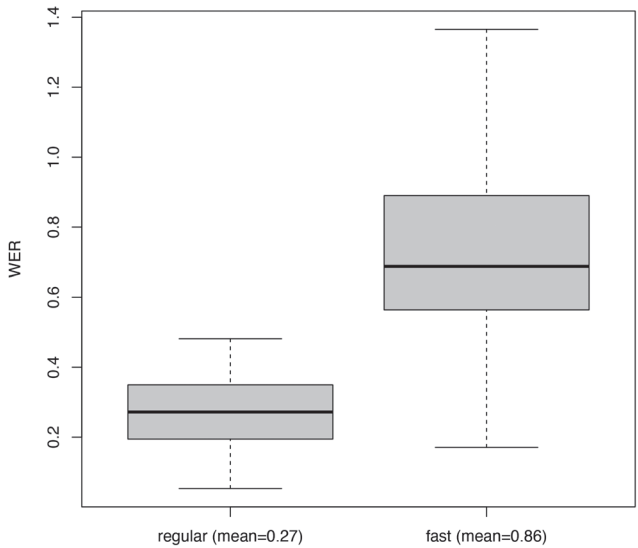

Figure 5. WERs in the regular and the fast-speaking conditions. Two outliers of the fast speech group (WERs = 4.21 and 2.32) were excluded from the figure.

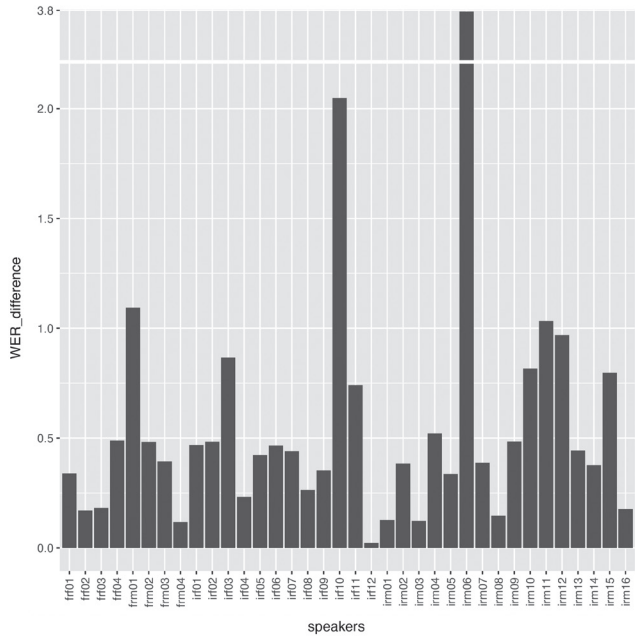

Figure 6. WER differences between fast and regular speech for each speaker. The differences are calculated by subtracting speaker-specific WERs of regular speech from those of fast speech.

Figure 5 reveals a drastic increase of WER in fast speech; whereas the mean WER is 0.27 in regular speech, in fast speech it increases to 0.86. The differences between the two speaking conditions for individual speakers are presented in Figure 6. This result shows that fast speech has a negative effect on ASR accuracy for all speakers, although the amount of difference varies between speakers.

To evaluate the relationship between changes in the prosodic features and speech intelligibility, six correlation tests were carried out using a Benjamini & Hochberg -adjusted significance level of 0.05. The results are presented in Table 2.

Table 2. Correlations between mean values of WER and the mean prosodic features of the speakers. The correlation tests included both speaking conditions. Statistically significant correlations are marked using bold type.

feature	*male speakers*		*female speakers*	
	r	*p*	*r*	*p*
s_1	**–0.48**	**0.004**	–0.31	0.088
SPI	**–0.46**	**0.006**	–0.30	0.102
f0 (Hz)	–0.29	0.088	–0.06	0.725
eb1kHz	**0.43**	**0.008**	**0.49**	**0.008**
duration (s)	**–0.57**	**0.001**	**–0.60**	**0.001**
AR	**0.62**	**>0.001**	**0.64**	**0.001**

The correlation between the s_1 and WERs was statistically significant for the male speakers (r=–0.48), but not for the female speakers (r=–0.31). Similarly, the correlation between the scalar mean SPIs and WERs was statistically significant for the male (r= –0.46) but not for the female speakers (r=–0.30). Therefore, the SPI-related correlations demonstrate at least partial association between decreased prominence and speech intelligibility.

The energy values, syllable durations and ARs also correlated with WERs, which shows that lower speech intelligibility is associated with higher energy proportion below 1 kHz at a faster speech tempo. The correlation was especially strong between AR and WERs (r=0.62 and 0.64), demonstrating the strong negative effect of speaking fast on speech intelligibility. However, there was no significant correlation between f0 and WERs. Overall, the results considering the relationship between prosodic features and WERs were largely similar for the male and the female speakers.

4. Discussion

In the Introduction, two hypotheses were presented: speaking fast (1) decreases prosodic prominence and (2) deteriorates speech intelligibility. The results confirmed both of them. The mean values of articulation rate, syllabic duration, f0, energy proportion

below 1 kHz and SPI revealed a significant change towards lower prominence when the speakers spoke fast compared to regular speech tempo.

The dynamic changes in prosodic prominence were investigated using fPCAs, which revealed the major modes of variation in the fSPIs. The changes between the two speaking conditions were first examined for the male and the female speaker groups and then for the individual speakers in the PC score spaces. The first PC and the s_1, which explained over 80% of prominence variation, were mainly related to the overall height of the mean fSPIs. The second PC and the s_2, which explained only 1.6% of the variation, were more associated with the timings of the peaks in the mean fSPIs. Even though the clusters of the two speaking conditions partly overlapped in the PC score spaces, each speaker's s_1 was systematically lower in fast speech, indicating lower fSPIs. The s_2 variation was found to be unrelated to the speaking conditions. Thus, the functional results mainly supported the findings from the conventional prosodic analyses, but also showed rather high inter-speaker variation in prosodic prominence. Moreover, they verified that prosodic prominence is consistently (dynamically) lower in fast speech, which would not have been possible using conventional statistics.

Finally, the negative effects of speaking fast on speech intelligibility were established; in terms of mean WERs, the ASR accuracy decreased drastically from 0.27 to 0.86 when the speakers spoke fast. Even though the amount of decrease in WER varied between the speakers, ASR accuracy decreased for every speaker during fast speech. In addition, there was a statistically significant correlation between the WERs and most of the studied prosodic features.

Overall, this study has shown that when speakers intentionally alter their articulation rate, this has a holistic effect on speech prosody. Hence, the results suggest that it might be difficult, or even impossible, for speakers to alter speech tempo without an impact on other prosodic features. One of the few exceptions according to the previous literature might be speech rhythm, which was shown to have no significant within-speaker variation in different tempo conditions (Dellwo et al., 2015). Nevertheless, if the implication above holds, different prosodic aspects of speech can be even more connected than has been assumed in previous studies. Therefore, an aim for future studies would be to verify whether or not speakers are capable of conducting only tempo-related changes in speech prosody. In order to achieve this aim, functional data analyses can provide an efficient methodological framework.

5. Conclusions

In this study, prosodic changes caused by an increase of speech tempo were investigated. Dynamic changes in prosodic prominence were studied using SPI, a novel prominence measure, and functional PCA. In addition, the effects of increased tempo on speech intelligibility were evaluated using ASR. The results confirmed an expected increase in articulation rate and decrease in syllable duration in fast speech. In addition, energy proportion below 1 kHz was found to increase and f0 and SPI to decrease. fPCA verified dynamic changes in the functional SPIs, showing a systematic decrease for each speaker in fast speech. Finally, automatic transcriptions using ASR substantiated the neg-

ative effect of speaking fast on speech intelligibility. In addition, most of the prosodic measures correlated with the ASR accuracy.

REFERENCES

Boersma, P. & Weenink, D. (2020). Praat: doing phonetics by computer [Computer program]. Version 6.1.32. url: http://www.praat.org.

Corretge, R. (2020). Praat Vocal Toolkit. url: http://www.praatvocaltoolkit.com.

Cronenberg, J., Gubian, M., Harrington, J., & Ruch, H. (2020). A dynamic model of the change from pre-to post-aspiration in Andalusian Spanish. *Journal of Phonetics*, 83, 1–22. doi: 10.1016/j.wocn.2020.101016.

Cummins, F., Grimaldi, M., Leonard, T., Simko, J. (2006). The chains corpus: Characterizing individual speakers. *Proceedings of SPECOM*, Citeseer, pp. 431–435.

De Jong, N. H. & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behaviour Research Methods, 41*, 385–390.

Dellwo, V., Leemann, A., & Kolly, M. J. (2015). Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors. *The Journal of the Acoustical Society of America*, 137(3), 1513–1528.

Greenberg, S., Carvey, H., Hitchcock, L., & Chang, S. (2003). Temporal properties of spontaneous speech – a syllable-centric perspective. *Journal of Phonetics*, 31, 465–485.

Gubian, M., Boves, L., & Cangemi, F. (2011). Joint analysis of f0 and speech rate with functional data analysis. *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* IEEE, Prague, Czech Republic, pp. 4972–4975.

Gubian, M., Cangemi, F., & Boves, L. (2010). Automatic and data driven pitch contour manipulation with functional data analysis. *Speech Prosody*, Chicago, IL, USA.

Gubian, M., Torreira, F., & Boves, L. (2015). Using functional data analysis for investigating multidimensional dynamic phonetic contrasts. *Journal of Phonetics*, 49, 16–40. doi:10.1016/j.wocn.2014.10.001.

Hazan, V. & Markham, D. (2004). Acoustic-phonetic correlates of talker intelligibility for adults and children. *The Journal of the Acoustical Society of America*, 116, 3108–3118.

Janse, E. (2004). Word perception in fast speech: artificially time-compressed vs. naturally produced fast speech. *Speech Communication*, 42, 155–173. doi: https://doi.org/10.1016/j.specom.2003.07.001.

Janse, E., Nooteboom, S., & Quené, H. (2003). Word-level intelligibility of time-compressed speech: prosodic and segmental factors. *Speech Communication*, 41, 287–301.

Kisler, T., Reichel, U., & Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language,* 45, 326–347. doi: http://dx.doi.org/10.1016/j.csl.2017.01.005.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the h&h theory. *Speech production and speech modelling*, Springer, pp. 403–439.

Mayo, C., Aubanel, V., & Cooke, M. (2012). Effect of prosodic changes on speech intelligibility. *Proceedings of INTERSPEECH*, Portland, USA September 9-13, pp. 1708–1711.

Niebuhr, O., & Kohler, K. J. (2011). Perception of phonetic detail in the identification of highly reduced words. *Journal of Phonetics*, 39(3), 319–329.

Patel, R. & Schell, K.W. (2008). The influence of linguistic content on the lombard effect. *Journal of Speech, Language, and Hearing Research*, 51(1), 209–220. doi: 10.1044/1092-4388(2008/016)

Ramsay, J.O., Hooker, G., & Graves, S. (2009). *Functional data analysis with R and MATLAB*. NY: Springer.

Reetz, H. (2009). *Phonetics: transcription, production, acoustics, and perception.* Oxford: Wiley-Blackwell.

Roettger, T.B., Winter, B., & Baayen, H. (2019). Emergent data analysis in phonetic sciences: Towards pluralism and reproducibility. *Journal of Phonetics,* 73, 1–7.

Stanton, B., Jamieson, L. & Allen, G. (1988). Acoustic-phonetic analysis of loud and lombard speech in simulated cockpit conditions. *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*, pp. 331–334. doi: 10.1109/ICASSP.1988.196583.

Tavi, L. & Werner, S. (2020). A phonetic case study on prosodic variability in suicidal emergency calls. *International Journal of Speech, Language & the Law*, 27, 59–74.

Tomashenko, N., Wang, X., Vincent, E., Patino, J., Srivastava, B. M. L., Noé, P.G., Nautsch, A., Evans, N., Yamagishi, J., O'Brien, B., & Chanclu, A. (2021). The voiceprivacy 2020 challenge: Results and findings. arXiv preprint arXiv:2109.00648.
Zellers, M., Gubian, M., & Post, B. (2010). Redescribing intonational categories with functional data analysis. *Proceedings of INTERSPEECH*, Makuhari, Japan, pp. 1141–1144.

Lauri Tavi
School of Humanities
University of Eastern Finland
Agora, Finland
E-mail: lauri.tavi@uef.fi

2022 ACTA UNIVERSITATIS CAROLINAE PAG. 65–81
PHILOLOGICA 1 / PHONETICA PRAGENSIA

VARIATION IN SPEECH TEMPO AND ITS RELATIONSHIP TO PROSODIC BOUNDARY OCCURRENCE IN TWO SPEECH GENRES

JAN VOLÍN

ABSTRACT
The present study focuses on two problems connected with speech tempo. First, earlier research has been prevalently concerned with central tendencies while variation was mostly perceived as an auxiliary result. We believe, however, that information about data dispersion is essential for proper modelling and experiment design in the field of temporal structure of speech. Therefore, the present study provides reference values for some of the tempo metrics of variation that pertain to (a) between-genre differences, (b) within-genre differences, (c) inter-speaker differences, and (d) intra-speaker differences. Second, we tested the claim that faster tempi lead to fewer prosodic breaks in spoken texts. This claim had been supported by studies where a respondent was asked to produce the same text at various rates. We, on the other hand, pose a question of the number of prosodic breaks in speakers who are fast or slow inherently. The material used in the study represents two genres: poetry reciting and news reading, and we obtained recordings from 24 speakers in each genre. Apart from providing the quantifications, the outcomes suggest, for example, that the predisposition of individual speakers to produce fast or slow tempi differs between the two genres. The fastest speakers in news reading were not necessarily the fastest in poetry reciting. This result points at specific behaviour in different situations and invites caution concerning the idea of hard-wired speaking stereotypes in individuals. Also, the correlation between speakers' rates and the number of phrases they produced was significant only in news reading, not in poetry reciting. This result was corroborated by greater variation in prosodic boundary placement in news reading. In addition, the results offer an insight into the relationship between articulation rate and speech rate, together with the comparison of measurements in syllables per second and phones per second. The latter can be of interest since Czech (the language of the material) belongs to languages with a complex syllabic structure.

Key words: articulation rate, news reading, poetry reciting, prosodic boundary, speech rate

1. Introduction

Research in speech tempo or durations of speech sounds has provided a rich pool of results during its relatively long tradition. Besides sheer scientific curiosity, the motiva-

https://doi.org/10.14712/24646830.2022.29

tion for past studies varied between, for instance, synthesis-by-rule concerns (O'Shaughnessy, 1984; Carlson & Granström, 1986; Campbell, 1992), forensic use (Johnson et al., 1984; Künzel, 1997; Jessen, 2007), or automated competence assessment (Lennon, 1990; Cucchiarini, Strik & Boves, 1997; Graham & Nolan, 2019). It has been clearly established, however, that despite certain universal tendencies, the temporal structure of each language has to be studied on its own (e.g., Barik, 1977; Grosjean, 1980; Trouvain & Möbius, 2014).

Temporal patterns in the Czech language were repeatedly examined in the past and individual studies offered quite significant insights, although from today's perspective, researchers usually worked with smaller samples of speakers or with stylistically limited material. Moreover, some of the studies were published in sources that are currently difficult to access. A thorough dedicated study dealing with Czech is still missing. An outstanding exception is the monograph by Dankovičová (2001) which comprises several meticulous studies and offers valuable quantitative descriptions.

The conceptual fixation of linguists on lexical contrast sometimes leads to small appreciation of the temporal dimension in the prosodic structure of languages. Occasionally, it is even viewed as some sort of an insubstantial variable. Port (1979: 46) uses a strikingly harsh phrase: "phonologically irrelevant factors such as speaking tempo" (sic!), but this is probably a reflection of the widely held view at that time that phonology is solely concerned with segmental phonemes. We, on the other hand, argue that if tempo is systematically used in conveying any component of the communicated meaning, then it must have its own phonology.

One of the reasons for underestimating the functions of tempo in speech is probably methodological: the research is relatively poorly equipped. Current analytical tools do not generate temporal tracks as readily as amplitude or F0 tracks. (Although for a simple but relatively crude method see Volín, 2009). An implicitly connected problem is the belief in the existence of the so-called 'personal tempo'. Palková (1994: 317) defines it as a mean speech-production rate typical of an individual speaker. Informal observations, indeed, lead to perceiving certain speakers as slow, while others as moderate or fast. This idea, again, has its roots in averaging across large speech materials and in disregard for local contextual variation.

We dare to assume that rather than a personal 'signature tempo', individuals display specific strategies when accelerating or decelerating their speech for specific communicative purposes. This was suggested, for instance, for English (Goldman-Eisler, 1961), for French (Fougeron & Jun, 1998), for German (Trouvain & Grice, 1999) or for Greek (Fourakis, 1986). All of these studies, however, follow the common experimental paradigm: various speech tempi are elicited on request. An individual speaker is asked to establish his/her 'normal' rate and relative to that produce a fast/slow or a very fast/very slow version of the same text. Therefore, the speakers' judgements put the productions into classes of rates, but their ideas of what is very fast or very slow might be quite disparate. Nevertheless, the change in an individual behaviour when switching between tempi provides important information about the production of various speech units.

One of the more recent examples of the above-presented paradigm is the study by Werner and colleagues who focused on silent pauses and their association with various tempi produced by a speaker (Werner et al., 2022). The relevance of this study to our

present goals is in that besides others, the authors also used recordings of Czech speakers. However, the authors were interested solely in silent pause modelling and they did not provide any exact quantification of the rates in their material.

In contrast with that, our present study targets two areas of interest: (1) providing exact variation values based on a larger sample of speakers, and (2) correlating the occurrence of prosodic boundaries for fast or slow speakers in their own comfortable modes. The latter means that our speakers did not modify their tempi upon request. Instead, as a group, they created a continuum from slow to fast through their unconscious planning of 'adequate' rate for the given genre. Two speech genres were examined (see below). With regard to variation, we aim at (a) between-genre differences, (b) within-genre differences, (c) inter-speaker differences, and (d) intra-speaker differences.

2. Method

2.1 Material

The two genres examined were *poetry reciting* (POR) and *news reading* (NWS). POR was represented by three Czech poems (P1, P2, P3) from the beginning of the 20th century. Each of them comprised 20 verse lines and in agreement with general conventions of that period they were rhyming. In these poems, consecutive pairs of verse lines were analysed as prosodic wholes (referred to as 'speech units' below) since the pairs also formed distinct semantic units. This was especially clear in the poem P2, which was published in two-line stanzas. The other two poems had four-line stanzas, but major punctuation marks were prevalently present at the end of the second and fourth line. There are indications that the speakers produced the poems with the reflection of this fact (whether conscious or unconscious). Each speaker produced 30 such verse pairs (3 × 10) comprising 584 syllables in total. The title and the pause after it were excluded. The titles were read in disparate ways and the first pause was manifestly longer than all other pauses within the text and reflected some sort of preparatory strategy of the speakers rather than the properties of the text. Quite a few poetry readers actually seemed to be 'bracing' themselves for the 'real' beginning after the title.

The genre of *news reading* (NWS) was represented by four paragraphs (news items) of a realistic news bulletin (NI1, NI2, NI3, NI4). The actual text originally comprised six paragraphs plus some introductory and concluding phrases, but these phrases together with the first and the last paragraphs were excluded from analyses in order to balance the extent of the material used. Even despite this measure, the NWS text still consisted of 700 syllables. In parallel to verse pairs in POR, the NWS was analysed in sentences. Each speaker produced 19 of those in the four paragraphs analysed. Given the disparate structuring of the POR and NWS material, the mean length of a verse pair in our material was 19.2 syllables while that of a sentence was 36.8 syllables.

All recordings were processed identically. Forced alignment for words and phones was performed with Prague Labeller (Pollák, Volín & Skarnitzl, 2007), manual corrections and further labelling were carried out in Praat (Boersma & Weenink, 2019). The data were extracted with dedicated Praat scripts.

Individual poems and news bulletin paragraphs will be referred to as *genre units*. These should not be confused with *speech units*, i.e., verse pairs in poems and sentences in news.

2.2 Speakers

There were 24 speakers (12 female + 12 male). All speakers were current or former university students majoring in philological programmes. Their mother tongue was Czech and their ages ranged from 20 to 32 years. They volunteered after they had read an advertisement calling for people with inclination to poetry and without speech disorders or hearing problems. Financial remuneration was offered. The recording procedure was almost the same for POR and NWS material (a single exception is described below). Speakers were given individual poems or news paragraphs (= genre units) on separate sheets of paper, and were asked to get familiar with the contents and form of each of them. They were allowed to practice individual parts of the texts for as long as they needed. Then they were asked to read out the poem or paragraph as if talking to audiences. To alleviate the situational stress, the speakers were reassured that any mistakes would be edited out and their performance would be strictly anonymous. They were also invited to self-correct, i.e., to read out any speech unit again if they were not satisfied with the outcome. All recordings were made in the sound-treated studio of the Institute of Phonetics in Prague. The only difference in the procedure was the fixed order of paragraphs in NWS (according to the original news bulletin) and random order with fillers in the case of poems in POR.

2.3 Measurements

There is an array of descriptive statistics that reflect central tendencies and variation in a data set. However, certain considerations limit their use in given cases. The current study deals predominantly with rates, hence, harmonic mean had to be used when averaging tempi across several units that belong together. Arithmetic mean, on the other hand, was used when tempo of a unit was its descriptor and variation among units needed to be captured. With regard to metrics of variation, we argue that given our current goals the most beneficial ones are the variation range and variation coefficient. Variation range (Rg_{var}) is the distance between the lowest (minimum) and the highest (maximum) value in the set. In literature, it is often presented just by these edge values, but we find it convenient to report the distance itself as well.

Variation coefficient (C_{var}) is the ratio between the arithmetic mean and the standard deviation from that mean expressed as a percentage. Unlike variation range above it does not depend on two values only, it is calculated with all the data points in a set. As a rule of thumb, coefficients below 30% are considered to represent concentrated data, while coefficients over 50% reflect high dispersion in the data (e.g., Skalská, 1992: 12).

The current presentation practice favours measurements both in syllables per second (syll/s) and phones per second (pho/s). Since the relationship between the two is not straightforward in languages with complex syllabic structures (Pfitzinger, 1998; Koreman, 2006), we will report both rate units.

Outcomes of statistical significance tests concerning differences found will be considered significant at the level of $\alpha = 0.05$, and so will the correlation coefficients. However,

approximate values of *p* will be provided, as customary in current empirical research reporting.

2.4 A terminological note

The term *speech tempo* will be used as a general term (hyperonym) covering other, more specific metrics. The plain term *tempo* will also refer to speech tempo in the present text. *Articulation rate* (AR) is conventionally calculated as number of speech units per unit of time with the exclusion of pauses, i.e., only articulation of lexical items is considered. *Speech rate* (SR), on the other hand, includes pauses into the calculation. It expresses a number of speech units produced per unit of time throughout all speech activity, that is with non-lexical items and pauses included. Logically, for the same stretch of spoken text speech rate cannot be higher than the articulation rate. If there are no pauses and other non-lexical elements, it must be equal, otherwise it is lower.

3. Results

The results concerning speech tempi and their variation will be presented in the following order: (1) the differences between genres, (2) differences among genre units, i.e., within-genre differences, (3) differences between speakers, i.e., inter-speaker differences, and (4) differences among speech units produced by a speaker, i.e., intra-speaker variation. Subsequently, Section 3.5 describes the relationship between the number of prosodic phrases produced and the speakers' tempi.

3.1 Between-genre differences

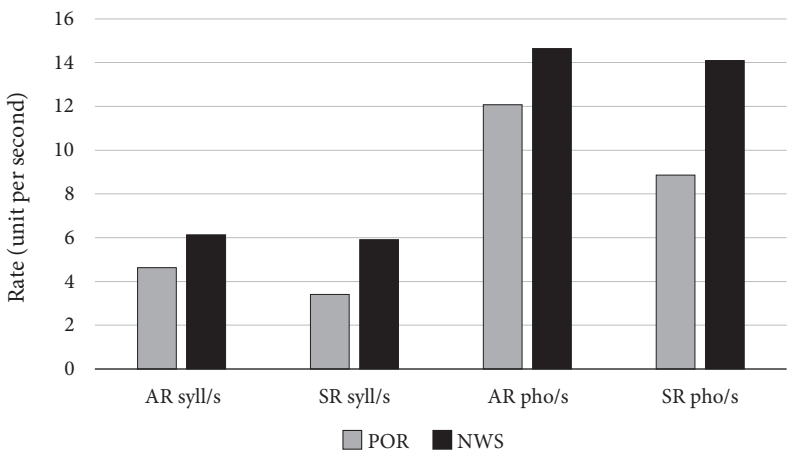

Figure 1. Mean articulation and speech rates (grand means) in two genres: poetry reciting (POR) and news reading (NWS). Values in syllables per second (syll/s) are on the left, phones per second (pho/s) on the right.

Mean articulation and speech rates between the two genres differed: the news reading (NWS) was on average always notably faster than poetry reciting (POR). Articulation rate was faster by 1.5 syll/s or 2.6 pho/s, while speech rate was faster by 2.5 syll/s or 5.24 pho/s. All the differences are displayed in Figure 1. They were tested by ANOVA for repeated measures, which returned highly significant results in all four cases: $F(1, 23) = 287.5$, $p < 0.001$; $F(1, 23) = 181.9$, $p < 0.001$; $F(1, 23) = 461.5$, $p < 0.001$; $F(1, 23) = 355.1$, $p < 0.001$ (arranged left to right after Fig. 1).

As to our key concern, variation, Table 1 summarizes the selected descriptors. It has to be pointed out that one data point in these calculations is a genre unit, i.e., one of the poems or one of the news bulletin paragraphs. The computation is then based on 72 + 96 data points (24 speakers × 3 poem or 4 news items). The coefficient of variation (C_{var}) in articulation rate was below 10%, which signals highly concentrated values. Speech rate C_{var} was somewhat higher but still did not exceed 15%. It is useful to note that while C_{var} in AR is roughly equal in both genres, the poetry reciting is more varied in terms of SR. Obviously, this is caused by unequal pausing strategies of individual speakers.

Interestingly, the variation range (Rg_{var}) exhibits an opposite pattern: the speech rate values are comparable, while articulation rate values are more dissimilar. It has to be pointed out, though, that Rg_{var} depends on two values only, which clearly disregards the situation in the rest of the data set. As a metric, Rg_{var} is often reported as a useful descriptor, but it has to be considered with caution.

Certain insight can be added by inspection of the minima and maxima themselves. There are two facts to be noted. First, it is apparent that the differences between the two genres are slightly greater in maxima than in minima. Second, the fact that NWS is on average faster is not caused solely by the maxima: both the lowest and the highest values are shifted upwards.

Table 1. Variation metrics across poetry reciting (POR) and news reading (NWS) given for articulation rate (AR) and speech rate (SR), both expressed in syllables per second (syll/s) and phones per second (pho/s).

	C_{var} (%)		Rg_{var}		*Max*		*Min*	
	POR	NWS	POR	NWS	POR	NWS	POR	NWS
AR-syll/s	8.1	9.0	1.8	2.5	5.5	7.5	3.8	5.0
SR-syll/s	12.3	9.3	2.1	2.5	4.5	7.3	2.4	4.8
AR-pho/s	7.0	7.9	4.1	4.8	14.2	17.6	10.1	12.8
SR-pho/s	11.3	8.1	5.2	5.1	11.7	17.2	6.5	12.0

3.2 Within-genre variation

Figure 2 shows that the mean tempi of the three POR genre units (i.e., three poems: P1, P2, and P3) were not equal. The strongest effect of Genre unit was returned by a one-way ANOVA for articulation rate in *syllables per second*: $F(2, 69) = 26.19$; $p < 0.001$, with post-hoc Tukey test confirming all three poems significantly different from each other.

The same effect for speech rate in syllables per second was weaker: $F(2, 69) = 13.79$; $p < 0.001$, with post-hoc Tukey test suggesting significant differences between P1 on the one hand, and P2 and P3 on the other. (The significance of the difference between P2 and P3 was no longer present.). The test criterion was slightly smaller when the unit of *phones per second* was used, but the result was still highly significant for articulation rate: $F(2, 69) = 11.62$; $p < 0.001$, with post-hoc inspection identifying P3 significantly different from P1 and P2. Finally, the weakest effect of Genre unit was produced for speech rate in phones per second: $F(2, 69) = 6.93$; $p \approx 0.001$. The post-hoc Tukey test found only the difference between P1 and P3 significant.

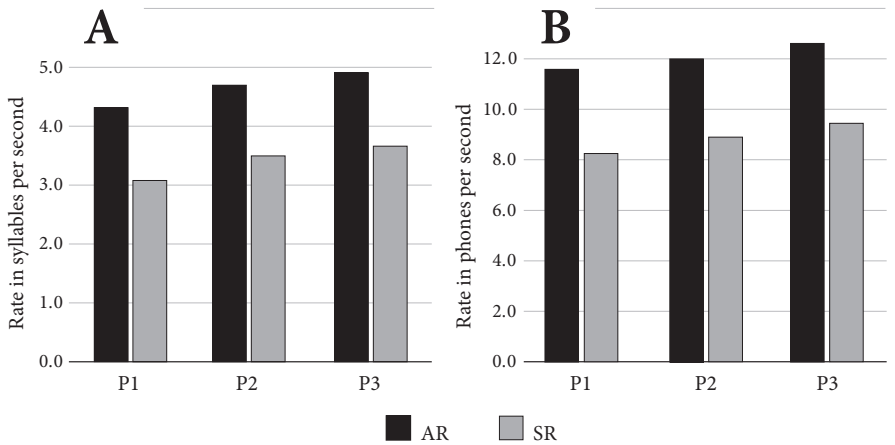

Figure 2. Mean tempi in the three investigated poems (P1, P2, P3). Panel A) captures the values in syllables per second, panel B) in phones per second. Darker columns represent articulation rate (AR), lighter columns pertain to speech rate (SR).

The same analysis was carried out for the NWS genre. Similarly to POR, Fig. 3 indicates that there were differences in the mean tempi of the individual genre units (i.e., the four bulletin paragraphs). It has to pointed out, that while poems were read in a random order with quite a lot of fillers in between, the news were read in a constant order dictated by the original broadcast. Thus, NI1 was always before NI2, etc. The figure shows how the mean tempo decelerates from the first domestic news through the second one and the foreign news down to the sports news with the lowest means.

The strongest effect of Genre unit was returned by a one-way ANOVA for AR in *syllables per second*: $F(3, 92) = 10.61$; $p < 0.001$. This is consistent with the test in POR reported above. The post-hoc Tukey test indicated NI1 significantly different from NI3 and NI4, and NI2 significantly different from NI4. The same effect for SR in syll/s was slightly weaker: $F(3, 92) = 10.06$; $p < 0.001$, but still highly significant. The post-hoc Tukey test suggested significant differences between NI1 and NI2 on the one hand, and NI3 plus NI4 on the other hand. The test criteria were smaller when the unit of pho/s was used, but the results were still significant both for AR and SR: $F(3, 92) = 4.79$; $p < 0.01$, and

$F(3, 92) = 3.74$; $p \approx .014$, respectively. The post-hoc test for the former found NI1 different from all the other NIs, whereas in the latter case significance was reached only for NI1 against NI3 and NI4.

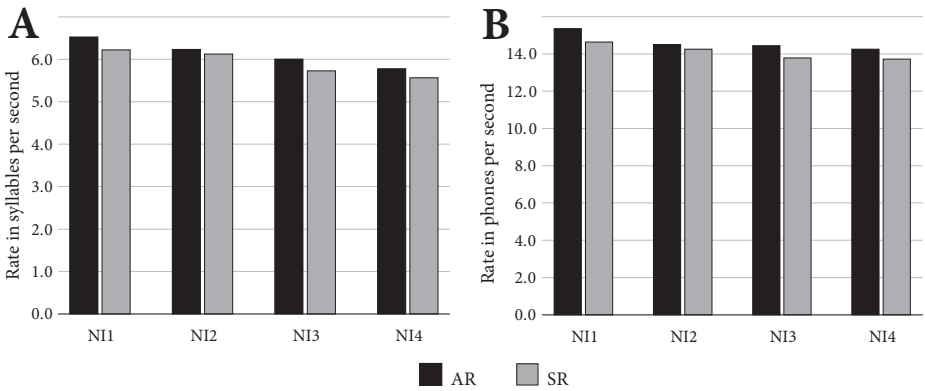

Figure 3. Mean tempi in the four investigated news items (NI1, NI2, NI3, NI4). Panel A captures the values in syllables per second, panel B in phones per second. Darker columns represent articulation rate (AR), the lighter ones represent speech rate (SR).

3.3 Interspeaker variation

The grand means across genres from Section 3.1 need to be broken into contributions by individual speakers. These are captured in Figures 4 and 5. The former displays the POR personal means, the latter the NWS means. The comparison of the figures confirms that the difference between AR and SR is smaller in news reading – a fact already noted in Section 3.1 above. It is also clear at first sight that the values produced by individual speakers are quite evenly distributed. There are no visible categorical breaks. Furthermore, it should be noted that the SR values are not exactly parallel to the AR values. This, again, indicates various pausing strategies among individuals. Also, the ordering individual rates by magnitude leads to roughly the same order in syll/s and pho/s – only small changes are observable.

The opposite is true when POR and NWS orderings are compared. Although in our current sample the slowest reciter is the slowest newsreader as well (speaker F10), the order of the other speakers by their tempi is not the same in POR as in NWS. This suggests that individuals have their specific inner concepts of each of the genres. In fact, only three speakers have the same position in the ordered set of POR and NWS. Seven speakers moved in the ordered data by one or two positions, four speakers moved by three or four positions. The remaining ten speakers moved by 5 or more positions, while four of those by even more than 10 positions.

Another way of looking at the same problem might be computation of Pearson's correlation coefficient between POR and NWS performances. This step returned $r = 0.44$ for

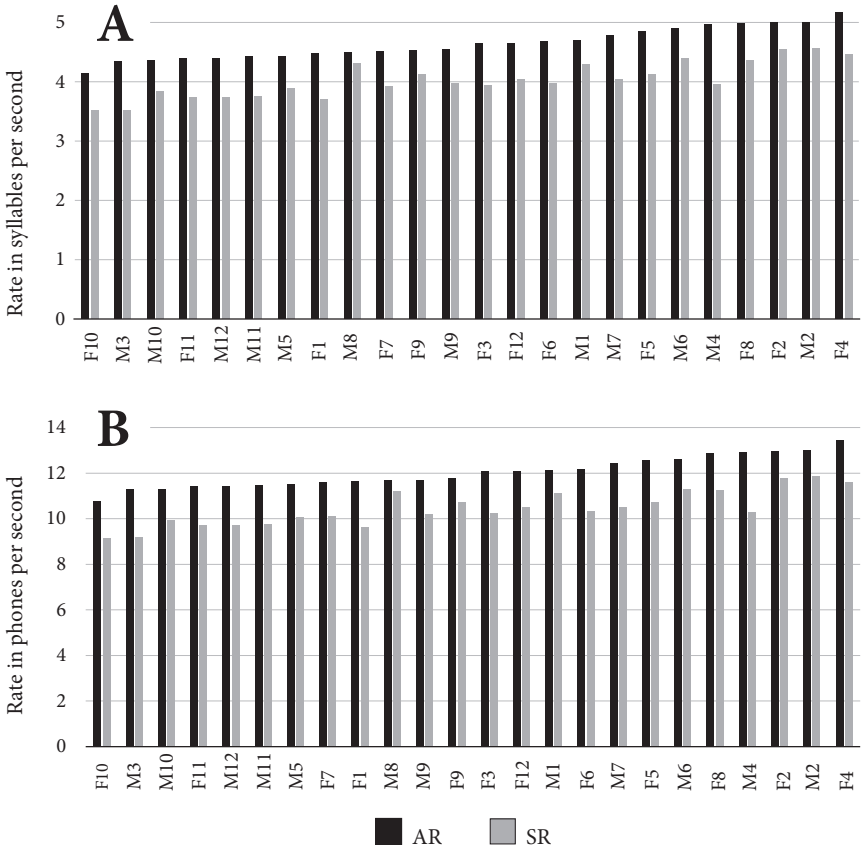

Figure 4. Mean tempi produced by individual speakers in poetry reciting (ordered by the AR values). Panel A) captures the values in syllables per second, panel B) in phones per second. Darker columns represent articulation rate (AR), lighter columns represent speech rate (SR).

both AR and SR in syll/s, and $r = 0.5$ for both AR and SR in pho/s (significant at the level of $\alpha = 0.05$). This suggests only moderate correspondence between the performances of a speaker in the two different speech genres.

Table 2 displays variation metrics across the sample of speakers. Unlike Table 1 above, Table 2 builds on individual people. Thus, for instance, *Min* refers to the slowest speaker, while Rg_{var} refers to the difference between the means of the fastest and slowest speaker under the given measurement condition.

It can be noted that the coefficient of variation (C_{var}) is below 8%, which means very low dispersion of the individual tempi. This is lower than the corresponding values in Table 1. The outcome is not surprising – Table 2 builds on mean tempi of individual speakers, while Table 1 reflects variation in mean tempi of individual genre units (poems or news paragraphs). The same holds for variation range (Rg_{var}): individual speakers differ less than individual genre units. For instance, the slowest and the fastest speakers in

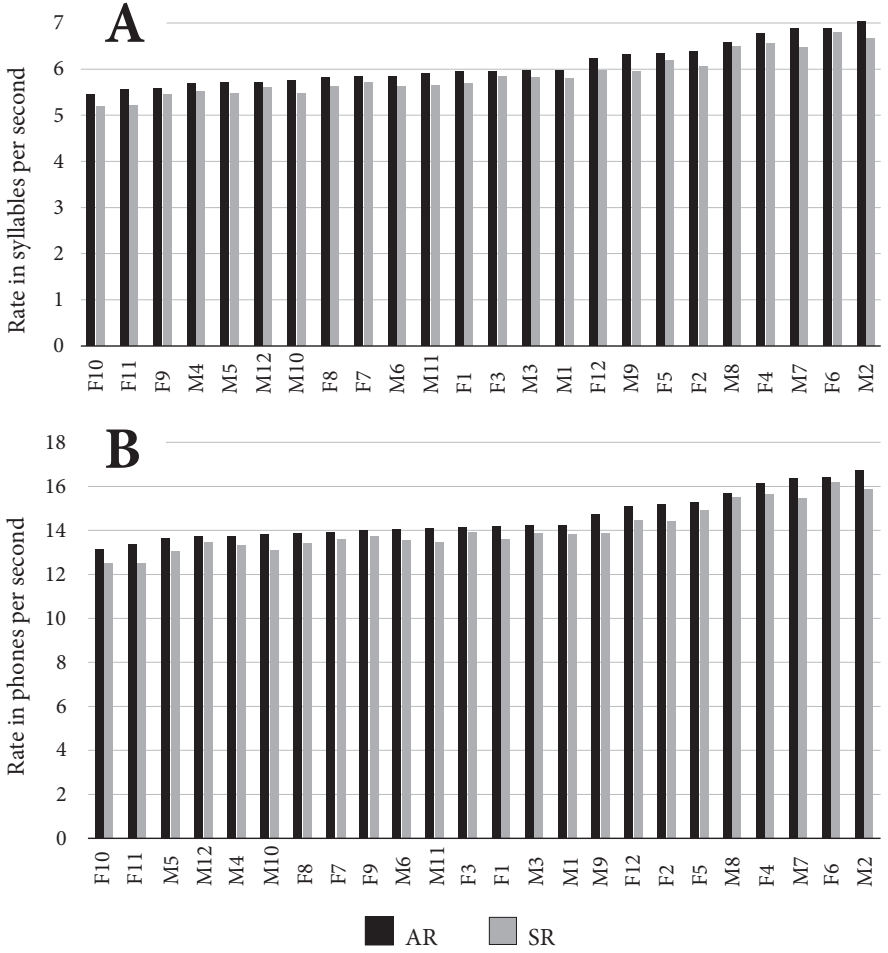

Figure 5. Mean tempi produced by individual speakers in news reading (ordered by the AR values). Panel A) captures the values in syllables per second, panel B) in phones per second. Darker columns represent articulation rate (AR), lighter columns represent speech rate (SR).

Table 2. Variation metrics across speakers in poetry reciting (POR) and news reading (NWS) given in articulation rate (AR) and speech rate (SR), both expressed in syllables per second (syll/s) and phones per second (pho/s).

	C_{var} (%)		Rg_{var}		*Max*		*Min*	
	POR	NWS	POR	NWS	POR	NWS	POR	NWS
AR-syll	5.7	7.5	1.0	1.6	5.1	7.0	4.1	5.4
SR-syll	7.6	7.7	1.1	1.6	4.6	6.8	3.5	5.2
AR-pho	5.7	7.2	2.7	3.7	13.5	16.8	10.8	13.1
SR-pho	7.5	7.4	2.7	3.7	11.9	16.2	9.2	12.5

POR differ by 1 syll/s, given that the fastest reciter spoke at AR of 5.1 syll/s while the slowest spoke at AR of 4.1 syll/s. Similarly, the fastest newsreader produced AR of 7.0 syll/s, while the slowest one 5.4 syll/s – hence the Rg_{var} of 1.6 syll/s.

All the minima in Table 2 (i.e., the slowest individuals) are unsurprisingly higher than the lowest values in Table 1 (i.e., the slowest genre units). It could be expected that, correspondingly, the maxima in Table 2 (i.e., the fastest individuals) would be lower than the maxima in genre units. However, this is only true for NWS and AR in POR. The speech rate in POR marginally diverges from this trend.

3.4 Intraspeaker variation

The variation of tempi produced by a single speaker (the within-speaker variation) can be illustrated by a histogram of values representing his or her speech units. Speaker M12 was identified as a typical individual with modal Rg_{var} since his Rg_{var} lay in the middle of the data set ordered by magnitude. His values are displayed in Figure 6.

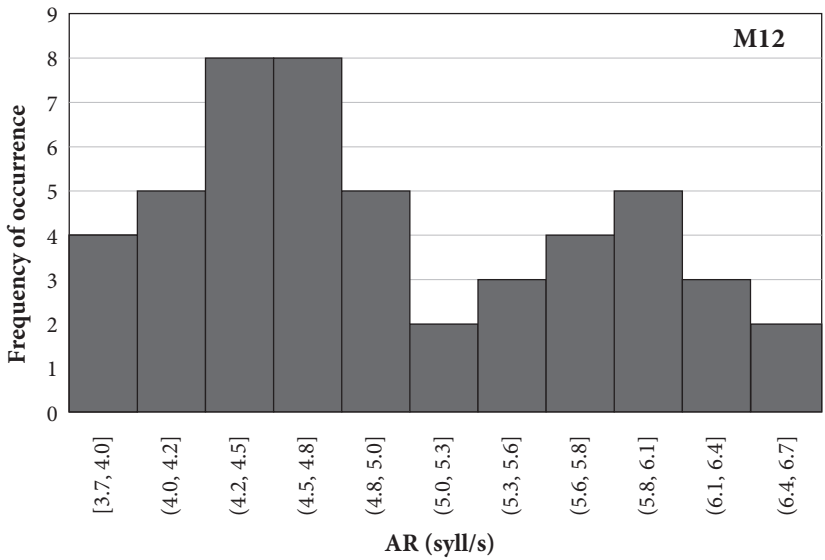

Figure 6. Histogram of AR values in speech units (n = 49) produced by speaker M12 (see text for selection reasons).

The first important fact to note is the bimodality of the histogram. Indeed, the articulation rates of POR were clearly lower than those of NWS, as signalled by highly significant effect of genre (Section 3.1). Thus, when collapsing data from two genres into one set, researchers map certain communicative potentials of a given speaker, but they should not necessarily expect normal (Gaussian) distribution of values within such a combined set.

The second detail to point out is the scale of intraspeaker variation, which is clearly larger than variation among means of individuals (analysed in the previous section). The difference between the slowest and fastest speech unit of this particular speaker was 3 syll/s.

Rather than mean values as in previous sections, we will present a few individual examples at this point to expose intraspeaker variation. (This is because the approach analogous to Sections 3.1 and 3.2 would require 24 tables of the Tab. 1 and Tab. 2 design, which would impair the lucidity of the presentation). The examples in Tables 3 and 4 were selected to represent the most monotonous, the most balanced, and the most varying speaker in each genre.

Table 3. Articulation rate metrics representing intraspeaker variation in three speakers of monotonous, balanced and changeable type in poetry reciting (POR) and news reading (NWS).

Genre	*Speaker*	*Min* (syll/s)	*Max* (syll/s)	C_{var} (%)	Rg_{var} (syll/s)
POR	monotonous	3.87	5.14	6.91	1.27
	balanced	3.68	5.30	9.30	1.62
	varying	3.40	5.69	10.95	2.29
NWS	monotonous	4.59	6.26	8.55	1.67
	balanced	4.21	6.65	9.69	2.44
	varying	5.49	9.87	15.27	4.37

Apart from the fact that all variation parameters are lower in POR than in NWS, it can be observed that the varying speaker in POR not only raises the maximum, but also lowers the minimum. This does not happen in NWS, although it is the same person. We might speculate that temporal strategies of an individual differ across speech genres. As to the other metrics, their values increase from monotonous to varying type. Analogous data for speech rate (SR) are displayed in Table 4.

Table 4. Speech rate metrics representing intraspeaker variation in three speakers of monotonous, balanced and changeable type in poetry reciting (POR) and news reading (NWS).

Genre	*Speaker*	*Min* (syll/s)	*Max* (syll/s)	C_{var} (%)	Rg_{var} (syll/s)
POR	monotonous	3.20	4.65	8.06	1.45
	balanced	3.08	5.21	11.42	2.13
	varying	2.92	5.69	14.82	2.77
NWS	monotonous	4.59	6.26	9.28	1.67
	balanced	4.98	7.65	10.48	2.67
	varying	4.34	9.46	17.60	5.12

Comparison between Tables 3 and 4 reveals that both C_{var} and Rg_{var} are higher in speech rate than in articulation rate. A similar trend was already reported in previous sec-

tions. Greater variation is obviously caused by the use of pauses, which lower the minima more than the maxima. For instance, the slowest speech unit of the monotonous speaker has AR that is 75.3% of her fastest unit. In terms of SR, it is only 68.8%.

For the sake of brevity, we will not report analogous results for measurements in pho/s. They were inspected and established as patterning consistently with the measurements in syll/s displayed in Tables 3 and 4.

A final observation presented in this section concerns an interesting difference in distribution of the variation metrics of C_{var} and Rg_{var}. Figure 7 documents that while the C_{var} values are spread more or less symmetrically and peaking at about the middle, the Rg_{var} values have massively skewed distribution with most data points in the low values and progressively fewer in high values.

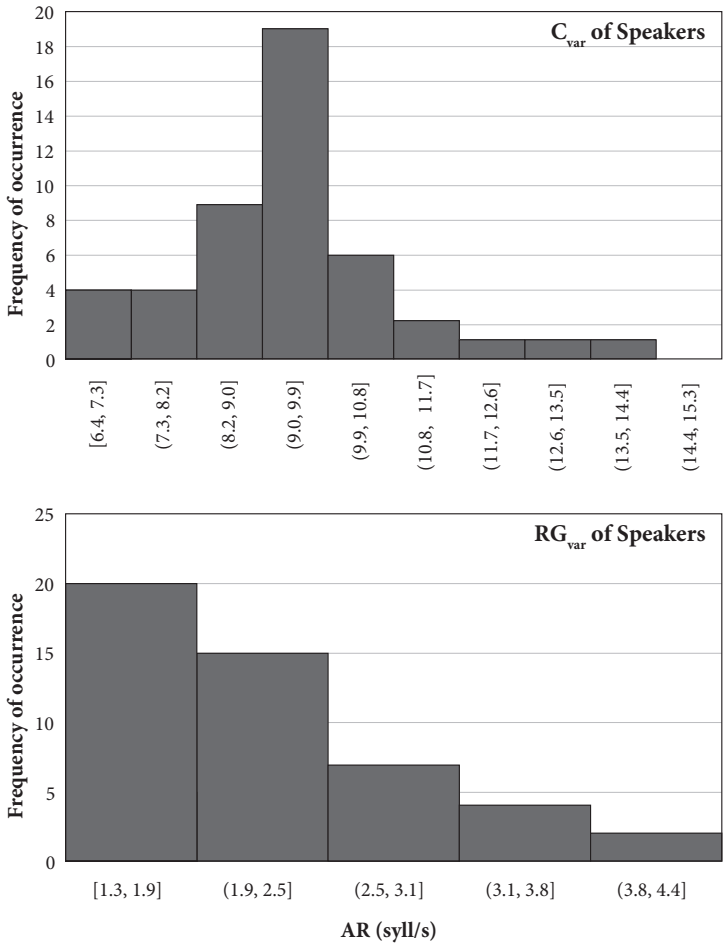

Figure 7. Histograms of within-speaker C_{var} and Rg_{var} values produced in individual performances. Measurement condition: AR in syll/s.

3.5 Division into prosodic phrases

The major question answered in this section concerns the frequency of occurrence of prosodic phrases in relation to AR or SR. Only full prosodic phrases were considered (i.e., intonation phrases in ToBI terminology). On average, the speakers produced 158 prosodic phrases each, of which 96 were in POR and 62 in NWS. The lowest number of prosodic phrases produced by one speaker was 131, while the highest number was 175. These extremes delimit the variation range and they were both produced by male speakers. (However, since male/female opposition was not examined in this study, this fact will not be elaborated on).

The declared focus of the present study is variation. The speakers produced exactly the same texts in two genres, but their production could differ by 44 prosodic boundaries. This span seems impressive, however, in terms of C_{var} it is only 7.6%, which indicates highly concentrated data. An overview for the sample of present genres is provided in Table 5. Interestingly, when the variation metrics are calculated for each genre separately, C_{var} emerges markedly higher for NWS than for POR (Table 5). This suggests that poem structuring guides the speakers more firmly, whereas the news texts provide greater freedom for prosodic boundary placement. Nevertheless, C_{var} of 12.2% still reflects concentrated data.

Table 5. Variation metrics for the number of prosodic phrases in the examined texts in poetry reciting (POR) and news reading (NWS). The metrics Rg_{var}, *Max*, *Min* are given in number of phrases.

	C_{var} (%)	Rg_{var}	*Max*	*Min*
POR	7.1	26	108	82
NSW	12.2	28	75	47
Both	7.6	44	131	175

When correlating speakers' speech rates with the number of prosodic phrases they produced (Pearson's formula), the coefficients were $r = -0.51$ for AR both in syll/s and pho/s, and $r = -0.64$ for SR both in syll/s and pho/s. This result applies to data undifferentiated for genres. When the numbers of prosodic phrases were split by genre, the significant correlation disappeared for POR, but strengthened for NWS, where the correlation coefficients were: $r = -0.58$ for AR in syll/s, $r = -0.67$ for SR in syll/s, $r = -0.57$ for AR in pho/s, and $r = -0.66$ for SR in pho/s.

4. Discussion

The two objectives set for the current study were met: (1) the variation of tempo in two speech genres was quantified, and (2) the relationship between the articulation rate/speech rate on the one hand, and the number of prosodic phrases produced in a text on the other hand, was examined.

As to the latter, our expectations were based on older laboratory experiments where the same speakers were asked to pronounce identical sentences in slow, moderate and fast rates, and their fast speech contained fewer phrases. In our study, we modified the research question and asked whether the speakers who use habitually faster or slower speech tempi would follow such a pattern as well. The results of correlation analyses showed that to some extent they do so. The returned coefficients were, indeed, negative, which means fewer prosodic breaks with faster rates. However, the relationship between the two variables does not seem to be very strong: only about 30% of variance could be explained when all our speech material was combined ($r^2 \approx 0.30$). What is even more interesting, though, is the difference between articulation rate and speech rate. The correlation coefficients were clearly higher for SR, suggesting that there is some systematicity in pausing, and that pure articulation is less flexible. Moreover, the statistical significance of the correlation coefficients was confirmed only for news reading.

This fact supports the increasingly prevalent claims that speech styles and genres matter in phonetic research (Wagner et al., 2015). The two genres examined in the present study differed in other aspects as well. Articulation rate in NWS was by 1.5 syll/s faster than in POR, and in terms of speech rate the difference was even larger: 2.5 syll/s. This implies that pauses in poetry reciting occupy greater space. This fact also caused greater C_{var} in speech rate in POR. On the other hand, with respect to the occurrence of prosodic phrase boundaries, greater variation was ascertained in NWS than in POR (as expressed by C_{var}). This indicates stronger demand on certain prosodic structuring in poetry and greater space to manoeuvre in news reading.

In future, however, not only the number, but also the actual placement of prosodic boundaries should be examined. Clearly, the linguistic specification of the positions with high or low concord among speakers would be of interest.

The analysis of inter-individual differences suggested that the relative tempo in the group is not the same across the two genres. The correlation coefficient between the performances of the speakers in POR and NWS was only moderate (cf. Section 3.3). It follows that individual temporal inclinations should not be over-estimated. Although informal experience points at the existence of habitually slow or fast speakers, generalizations across speaking genres might be injudicious. While a few speakers might not differentiate between the genres by tempo, the majority seem to exhibit specific personal concepts of the genre temporal form.

On the other hand, the differences between speakers within a genre were surprisingly low. The coefficient of variation was below 8% in all four measurement modes. This might suggest that just as we share the lexicon and syntax of a language, we also share the *prosodic grammar* for various communicative purposes.

Unsurprisingly, the within-speaker variation turned out to be greater than variation based on large averaging. There were speakers whose performance could be classified as varied, while others could be labelled as monotonous. The varied performance meant C_{var} up to 17%, whereas the monotonous one would produce C_{var} below 10%. Again, the coefficients of variation in individual speakers were lower for AR than for SR, suggesting that individuals are more stable in their speed of articulation than in pausing. This fact invites a more thorough research into pausing strategies (as pauses were the only difference in calculations of AR and SR). On the whole, however, the results are in line with the

findings of Dankovičová (2001), who focused on changes in AR within prosodic phrases. She reported variation of about 10% and only exceptionally, mainly in phrase final positions, slightly over 15%. Similar results were implied by Goldman-Eisler (1961), even if the methodology does not allow for direct comparison.

Finally, it has to be stressed that the reference values which we have provided in the present study do not speak for the Czech population as a whole. The sample comprised young university-educated and philology-oriented people, who represent a sector of population with high level of literacy and relatively advanced language competences. For future research, expansion to other social groups of Czech-speaking population would be desirable. Likewise, various other speech genres should be mapped and contrasted with the present results. We believe that the topic of tempo variation should be pursued further with the aim to provide a solid basis for 'temporal phonology'.

5. Acknowledgement

The study was carried out with the support of GAČR (Czech Science Foundation), Project 21-14758S. The author also wishes to express his thanks to doc. Radek Skarnitzl for valuable comments on the draft of the study, and to reviewers for their thorough work.

REFERENCES

Barik, H. C. (1977). Cross-linguistic study of temporal characteristics of different types of speech materials. *Language and Speech, 20*, 116–126.

Boersma, P., & Weenink, D. (2019). *Praat: doing phonetics by computer*. Computer programme downloaded at http://www.praat. org/.

Campbell, W.N. (1992). Syllable-based segmental duration. In: G. Bailly, C. Benoit, T. Sawallis (Eds.) *Talking Machines. Theories, Models, and Designs*. Amsterdam: Elsevier Science Publishers. 211–224.

Carlson, R., & Granström, B. (1986). A search for durational rules in a real-speech database, *Phonetica 43*, 140–154.

Cucchiarini, C., Strik, H., Boves, L. (1997). Automatic evaluation of Dutch pronunciation by using speech recognition technology. In: S. Furui, B.H. Juang, W. Chou, (Eds.), *Proceedings IEEE Workshop on Automatic Speech Recognition and Understanding*, Santa Barbara, pp. 622–629.

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

Fougeron, C. & Jun, S.-A. (1998). Rate effects on French intonation: prosodic organization and phonetic realization. *Journal of Phonetics 26*, 45–69.

Fourakis M. (1986). An acoustic study of the effects of tempo and stress on segmental intervals in Modern Greek. *Phonetica, 43(4),* 172–188. https://doi.org/10.1159/000261769

Goldman-Eisler, F. (1961), The significance of changes in the rate of articulation. *Language and Speech, 4*, 171–174

Graham, C., & Nolan, F. (2019). Articulation rate as a metric in spoken language assessment. In: *Proceedings of INTERSPEECH*, Graz: ISCA pp. 3564–3568.

Grosjean, F. (1980). Comparative studies of temporal variables in spoken and sign languages: A short review. In: W. Dechert & M. Raupach (Eds.) *Temporal variables in speech*, pp. 307–312. Berlin: De Gruyter Mouton.

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

Johnson, C. C., Hollien, H. & Hicks, J. W. (1984). Speaker identification utilizing selected temporal speech features. *Journal of Phonetics, 12*, 319–326.

Künzel, H. J. (1997). Some general phonetic and forensic aspects of speaking tempo. *Forensic Linguistics, 4*, 48–83.
Koreman, J. (2006). Perceived speech rate: The effects of articulation rate and speaking style in spontaneous speech. *Journal of the Acoustical Society of America, 119*, 582–596.
Lennon, P. A. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning, 40*, 387–417.
O'Shaughnessy, D. (1984). A multispeaker analysis of durations in read French paragraphs. *Journal of the Acoustical Society of America, 76*, 1664–1672.
Palková, Z. (1994). *Fonetika a fonologie češtiny*. Praha: Karolinum.
Pfitzinger, H.R. (1998). Local speech rate as a combination of syllable and phone rate. In: *Proceedings of ICSLP '98 – Sydney*, vol. 3, pp. 1087–1090.
Pollák, P., Volín, J. & Skarnitzl, R. (2007). HMM-based phonetic segmentation in Praat environment. In: *Proceedings of XII*th *Speech and Computer* – SPECOM 2007, pp. 537–541.
Port, R.F. (1979). The influence of tempo on stop closure duration as a cue for voicing and place. *Journal of Phonetics, 7*, 45–56.
Skalská, H. (1992). *Úvod do biostatistiky*. Hradec Králové: LF UK.
Trouvain, J. & Grice, M. (1999). The effect of tempo on prosodic structure. *Proc. of the 14th International Congress of Phonetic Sciences*. San Francisco: IPA, pp. 1067–1070.
Trouvain, J. & Möbius, B. (2014). Sources of variation of articulation rate in native and non-native speech: comparisons of French and German. *Proceedings of 7th Speech Prosody,* Dublin, pp. 275–279.
Volín, J. (2009). Metric warping in Czech newsreading. In: R. Vích (Ed.) *Speech Processing – 19*th *Czech-German Workshop*, pp. 52–55. Praha: AVČR.
Wagner, P. Trouvain, J. & Zimmerer, F. (2015). In defense of stylistic diversity in speech research. *Journal of Phonetics 48*, 1–12.
Werner, R. Trouvain, J., & Möbius, B. (2022). Optionality and variability of speech pauses in read speech across languages and rates. In: S. Frota, M. Cruz, & M. Vigário (Eds.) *Proceedings of 11th Speech Prosody*, Lisbon, pp. 312–316.

Jan Volín
Institute of Phonetics
Faculty of Arts, Charles University
Prague, Czech Republic
E-mail: jan.volin@ff.cuni.cz

2022 ACTA UNIVERSITATIS CAROLINAE PAG. 83–95
PHILOLOGICA 1 / PHONETICA PRAGENSIA

INTRA- AND INTER-SPEAKER VARIABILITY OF VOWEL SPACE USING THREE DIFFERENT FORMANT EXTRACTION METHODS

ALŽBĚTA HOUZAR AND RADEK SKARNITZL

ABSTRACT

Individual speakers' voices display various unique patterns, one of the most prominent of which is vowel articulation. This study focuses on vowel space properties of 15 Czech speakers in read and spontaneous speech, comparing outputs of three formant extraction methods, measuring formants: (1) in the vowels' temporal midpoints, (2) as their mean from the vowels' middle thirds, and (3) in the vowels' articulatory targets. The results show extensive variability across speakers, but also great within-speaker variability between the two speech styles, with spontaneous speech manifesting more centralised vowel pronunciation than read utterances. The first two measurement methods did not yield systematically different results, while formant values extracted from acoustically defined articulatory targets lead to noticeably larger vowel spaces. The results suggest that care should be taken when interpreting formant values obtained by different methods.

Key words: vowel space area, vowel formants, intra-speaker variability, inter-speaker variability, Czech

1. Introduction

Variability is an inherent characteristic of human speech and in speech science, perhaps the best-known example is illustrations of speakers' vowel systems. While traditional depictions of a vowel system will show discrete points corresponding to individual phonological vocalic qualities, nothing could be further from phonetic reality of everyday speech.

This study deals with vowel formants, i.e., the resonance frequencies of the vocal tract. The lowest two resonances – F1 and F2 – and partly also F3 depend on the vowel quality (i.e., the momentary vocal tract setting), while higher formants tend to remain relatively stable (Reetz & Jongman, 2009: 184). F1–F3 values therefore significantly vary within an individual's speech, but their patterns to some extent reflect their idiosyncrasies and they differ across speakers.

https://doi.org/10.14712/24646830.2022.30

1.1 Formant-based parameters

Vowel formants can be parametrized using several methods. Among the most common ones is extracting formant values in individual vowels, with one vowel token characterized by a single value per formant. Extracting formant values from vocalic segments (tokens) of a voice sample allows for observing their variability in the given vowel (type) as well as across vowels. Vowel formants can be examined individually, but it is also possible to observe multiple formants at once, viewing an analyzed vowel segment as a point in a multi-dimensional space defined by the given formants. For example, by plotting individual vowels in a two-dimensional F1~F2 space, we obtain the speaker's vowel space. Such a plot correlates with the speaker's vocal tract physiology, but also their articulation habits: centralized articulation (hypoarticulation) yields a smaller vowel space, while more distinct vocalic articulation (hyperarticulation) results in an expanded space. A parameter based on vowel space that can be measured is vowel space area (VSA), i.e., its area expressed as the formant measurement unit squared. VSA is usually delimited by formant values of the most peripheral (that is, front/back/open/close) vowels, of which at least three are needed for VSA analysis (Fletcher et al., 2015) but it is possible to include other ones as well (as seen for example in Weirich & Simpson, 2013).

Another example of vowel formant parametrization is long-term formant distributions (LTFs). This metric was introduced by Nolan and Grigoras (2005), and it is determined by each formant's distribution throughout a voice sample regardless of individual vowel qualities. Formant values are extracted from equidistant points within vowel or voiced intervals, i.e., multiple formant values are extracted from one segment. LTFs reflect the dimensions of the speaker's vocal tract as well as their articulation habits such as a tendency towards palatalization or lip rounding (Nolan & Grigoras, 2005). Analogously to individual vowels' formants and vowel space, a multi-dimensional representation of LTFs (LTF1 and LTF2) is possible as well, resulting in vowel space density (VSD; Story & Bunton, 2017). F1 and F2 values extracted at multiple time points throughout a voice sample are plotted in a two-dimensional space defined by F1 and F2, while the points' density constitutes the third dimension. Similarly to vowel space and VSA, these metrics also reflect the speaker's vowel articulation patterns.

1.2 Vowel formant and VSA measurement methods

Formant values characterizing individual vowels may be obtained using several different procedures. The most straightforward one appears to be measuring formant values in the temporal midpoint of each vowel; this method appears to be used in the majority of current studies (e.g., Nolan & Grigoras, 2005; Skarnitzl et al., 2015; Pettinato et al., 2016; Cavalcanti et al., 2021).

Some authors argue that the temporal middle of the vowel may not be the optimal point for formant extraction. Jacewicz et al. (2007) measured formant values at 20% and 35% of the vowels' duration – as the authors state, "[t]hese two measurement locations present the most expanded characterization of the working vowel space", while formant values in later points "tend to portray relatively centralized vowels which would tend to reduce

the vowel space area" (Jacewicz et al., 2007: 1466). Fletcher et al. (2015) also presumed the articulatory target can be reached at a different point than in the vowel's temporal midpoint and examined formant values also in the articulatory target, i.e., "at a time where there was minimal movement in formant tracks – for the best approximation of the vowels' steady-state target" (Fletcher et al., 2015: 2134) between 20% and 80% of the vowel's duration. The results of their experiment indeed show that VSAs based on formant values extracted from the temporal midpoints and articulatory targets significantly differ, the latter being larger, which supports their hypothesis. Measuring vowel formants in the (automatically identified) articulatory targets[2] was also performed for example by Fletcher et al. (2017). In the studies mentioned above, F1 and F2 were measured at the same temporal point; however, as Rose (2015) points out, finding a single articulatory target in the vowel that would be universal for every formant can be problematic, as "the putative target lies at different duration points for each formant" (Rose, 2015: 4822); therefore, finding the target position of each formant separately could also be beneficial. To its advantage, this method, unlike the temporal midpoint (see above) or a mean of multiple values in the mid-section of the vowel (see below), is relatively independent of the placement of the vowel start- and end-points which can be inconsistent among labellers; as Fuchs (2017) points out, speech signal is a continuum, where finding distinct points inherently optimal for extraction of the given values can be problematic (Fuchs, 2017: 11). On the other hand, it can be presumed that an algorithm made to extract the formant's most extreme values might tend to cling to outliers.

The downside of extracting formant values from a single timepoint lies in the possible occurrence of erroneous values; an individual point may not be representative of the whole vowel, as it can be affected by a momentary fluctuation. Therefore, it may be more beneficial to use a mean value of several points within the vowel, excluding its edges that can be influenced by segmental environment. For example, Skarnitzl and Volín (2012) calculated F1 and F2 as an arithmetic mean of seven equidistant points in the middle third of a given vowel, while Tykalová et al. (2021) determined formants' values from "30-ms segment close to the middle section of a vowel where F1 and F2 formant patterns were visible and stable" (Tykalová et al., 2021: 931.e25).

Vowel space area measurements are also affected by the vowel selection which is employed. Fletcher et al. (2015, 2017) extracted F1 and F2 in three most extreme vowels – front [iː], open [ɐː], and back [oː] – in New Zealand English. Using three vowels was also opted for by Pettinato et al. (2016), who measured F1 and F2 of [iː], [ɔː] and [æ] in recordings of Southern British English speakers; as the authors say, those vowels were selected for analysis because "they were the most frequent per individual participant recordings" and "they had the best differentiation in terms of front–back and high–low distinctions and therefore covered the largest distances in the F1–F2 space" (Pettinato et al., 2016: 5). Three-vowel VSA was also analysed by Tykalová et al. (2021), using Czech phonologically short corner vowels [a], [ɪ], and [u] (although the long [iː] has a markedly more peripheral, "corner" quality in Czech). Jacewicz et al. (2007) analysed VSA based on four and five vowel qualities in three regional varieties of American English: [i], [æ], [ɑ]

[2] In this study, the "articulatory target" is defined acoustically as a presumed target of the articulatory gesture derived from vowel formant dynamics, analogically to the studies mentioned above.

and [u], subsequently also adding the diphthong [ɔɪ]. Simpson & Ericsdotter (2007) as well as Weirich & Simpson (2013) measured VSA in German using five vowels, namely [iː], [ɛ], [aː], [ɔ], and [uː].

Studies analyzing vowel formants also differ in their position on the scale between automatic formant extraction and manual measurements. Fully automatic formant extraction (applied, for example, by Weirich & Simpson, 2013 or Pettinato et al., 2016) can be considered unbiased and it is considerably faster; it is, however, prone to errors such as merged or missing formants (Tykalová et al., 2021: 931.e25). Potential errors can be avoided by manually correcting the extracted values (see, e.g., Fletcher et al., 2015 or Tykalová et al., 2021); the drawback of this approach lies in it being relatively time-consuming and, to some degree, subjective, potentially introducing the researcher's confirmation bias into the data.

This study analyzes F1 and F2 values in Czech monophthongs and focuses on vowel space. Its goal is to examine the three methods of vowel formant extraction which were described above. It appears that formant extraction from the middle third of vowel segments is most ecologically valid; the articulatory target method seems prone to extreme values, and extracting from a single temporal point in the middle of vowels increases the likelihood of obtaining erroneous values. We decided to apply all three methods to examine differences between their outputs and, by extension, comparability of studies using different formant extraction methods. Based on the formant values obtained, we will compare the variability of vowel space across speakers and speech styles.

2. Method

2.1 Material

Recordings from 15 Czech male speakers were used for the analysis. The speakers were randomly chosen from the Database of Common Czech (Skarnitzl & Vaňková, 2017) – a reference database for forensic purposes, containing voice samples from 100 male speakers aged between 19 and 50 (mean = 25.6 years, SD = 6.7 years), who performed several speaking tasks, representing different speech styles. For this study, recordings of two speech styles were used: (1) reading a phonetically rich text of 150 words (corresponding to reading time around 1 minute) and (2) a one-minute excerpt from a spontaneous interview where the speakers were encouraged to talk on a topic of their own choice. The recordings were obtained in quiet environments in the speakers' home or workplace (subtle acoustic discrepancy among individual speakers' recordings thus cannot be ruled out) in a WAV format with 48-kHz sampling frequency, using a professional portable recorder Edirol HR-09.

The recordings were automatically segmented using the Prague Labeller (Pollák et al., 2007); afterwards, phone boundaries were manually corrected in Praat (Boersma & Weenink, 2015), following segmentation principles described in Machač and Skarnitzl (2009).

The Czech phonemic inventory contains 10 monophthongs and 3 diphthongs: /iː ɪ ɛ ɛː a aː o oː u uː/ and /a͡ʊ̯ o͡ʊ̯ ɛ͡u̯/. In our analysis, only monophthongs were used. Since short

and long vowels' realizations generally do not significantly differ in their quality, the short vowels and their long counterparts were merged into single categories, with the exception of /ɪ/ and /iː/ (see Skarnitzl and Volín, 2012 or Šimáčková et al., 2012 for more details on the Czech vowel inventory); therefore, these two phonemes were treated as separate vowel qualities in the analyses below.

2.2 Extraction of formant values

F1 and F2 values in Hz were automatically extracted from each vowel token using three approaches:

- in a single timepoint in the middle of the vowel duration;
- as the mean value in the middle third of the vowel;
- from articulatory targets that were automatically detected based on the formants' shifts inside the vowel between 20% and 80% of the vowel's duration, the onset and offset 20% being excluded to eliminate potential interference of segmental environment. In front vowels /ɪ iː ɛ ɛː/, the articulatory target was identified as the point of F2 peak, in back vowels /u uː o oː/ as the point of F2 minimum, and in the open vowels /a aː/ as the point of F1 peak (analogously to Fletcher et al., 2015). Both F1 and F2 were measured at the described timepoints (i.e., F1 and F2 values for each vowel were extracted from the same timepoint).

This study's methodology represents a synthesis of formerly used procedures (described above), and its objective is to compare their output. The most prevalent of the examined methods appears to be formant extraction from the vowel midpoint which has been used by a variety of studies (see section 1.2). Extraction of mean formant values from the middle third of a vowel was employed by Skarnitzl and Volín (2012). The last method we analysed, i.e., extraction of formant values from the (acoustic) articulatory target, was based on the methodology described in Fletcher et al., 2015, who explain the procedure like this:

> Articulatory point measurement criteria were designed with the aim of extracting values at a time where there was minimal movement in formant tracks – for the best approximation of the vowels' steady-state target. For the front vowel, [iː], this point was set at peak F2 frequency; for the open [ɐː] vowel the target was extracted when F1 was at its maximum; and for the back [oː] vowel the target point was taken when the lowest value of F2 was reached (Watson and Harrington, 1999; Watson et al., 1998). (*Fletcher et al., 2015: 2134*)

All formant extractions were performed using the Burg LPC algorithm in Praat in three different settings. The default setting contained the detection of 5 formants between 0 and 5,500 Hz (i.e., 500 Hz more than a five-formant default range for an adult male – this expanded range was applied to avoid potential misses of higher formant values in front vowels). The secondary setting, on the other hand, used a reduced frequency range, detecting 5 formants within 0–3,000 Hz, in order to resolve potential errors of the first setting which tended to merge F1 and F2 in back vowels into one detected formant. Lastly, a tertiary setting was also present, extracting 10 formant values in 0–3,000 Hz band, in case neither one of the previous settings yielded accurate results.

All the extracted formant values were manually checked and corrected if necessary; in cases where the values obtained by the default setting prominently diverged from standard formant values for the given vowel quality, the spectrograms were both visually and auditorily inspected and when appropriate, the default values were replaced by those obtained using the secondary or tertiary settings (when those values included detection of random noise as formants due to the reduced frequency range, they were excluded based on the spectrogram visual inspection). Those abnormal values included F1 below 200 Hz or above 800 Hz, F2 below 600 Hz and above 1,500 Hz in back vowels, and F2 below 1,000 Hz and above 2,300 Hz in front vowels. Solely the strongly significant abnormalities in the default formant extraction output were manually corrected in order to avoid introducing confirmation bias into the data.

The output of this procedure was F1 and F2 values from 15 speakers, 2 speech styles, 6 vowel qualities and 3 extraction areas (temporal midpoint, middle third and articulatory target); in total, the final dataset contains F1 and F2 values of 24,646 vowels. The formant values were converted to the Bark frequency scale which is more psychoacoustically relevant compared to Hz.

2.3 Analysis

The vowel space area (VSA) was calculated for each speaker, each speech style, and each formant extraction method as the surface area in a two-dimensional F1~F2 space delimited by formant value medians of the individual vowel qualities, using the formula:

$$VSA = \left| \frac{(x_1 y_2 - y_1 x_2) + (x_2 y_3 - y_2 x_3) \ldots + (x_6 y_1 - y_6 x_1)}{2} \right|$$

with x being median F1, y being median F2 and numbers 1-6 corresponding to individual vowel qualities in the order [iː ɪ ɛ a o u]. VSA will be expressed in Bark squared (Bark^2).

Vowel space illustrations were prepared in R (R Core Team, 2021) and the *ggplot2* package (Wickham, 2016).

3. Results and discussion

The general results are depicted in Figure 1, which shows all the speakers' F1~F2 vowel space area in read and spontaneous speech, as extracted by the three methods described in section 2.2: using the vowels' temporal midpoint, the mean from the middle third of the vowel, and the articulatory target.

First, it is clear that speaking style affects VSA to a great extent: in most speakers, VSA in read speech (shown in circles in Fig. 1) is larger than in spontaneous speech (triangles); the difference is particularly salient in speakers HROK and especially KALT. Five speakers manifest an opposite tendency in at least one extraction method; only speaker NVAT's vowel space area turned out to be larger in spontaneous speech using all three extraction methods.

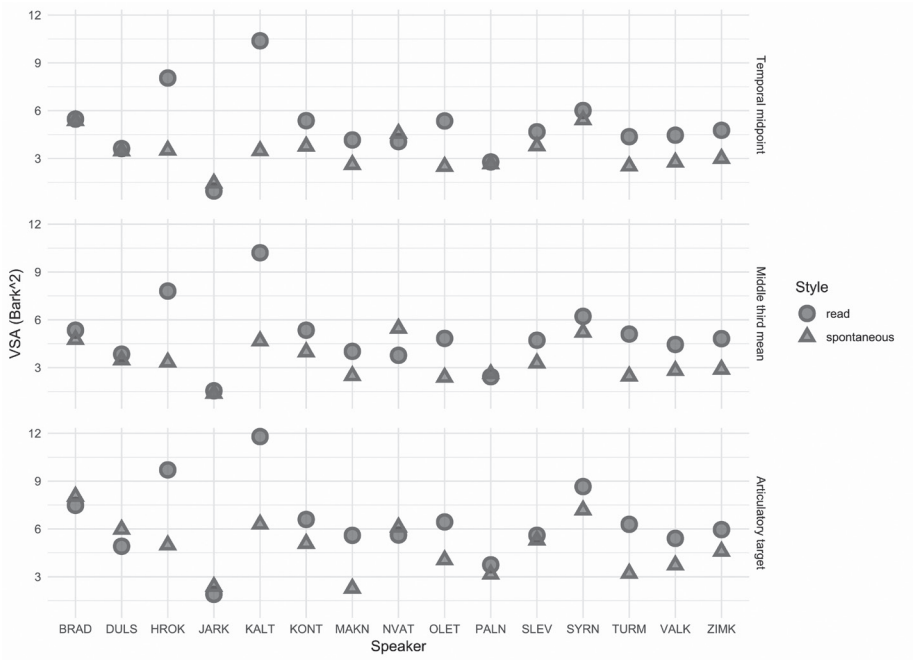

Figure 1. Vowel space area (VSA) of individual speakers in read and spontaneous speech, using three methods of formant extraction.

Second, although most speakers' VSAs fall between approximately 3 and 6 Bark^2, Figure 1 suggests that there are some between-speaker differences. Speaker JARK's vowel space area is strikingly small (see also below), and consistently so across the three measurement methods and across the two speaking styles. Indeed, his speech does sound remarkably centralized.

Third, the extraction methods themselves yield different results (more details follow below). It is not surprising that, almost without exception, the largest VSA is obtained by the articulatory target method, shown in the bottom panel of Figure 1 (*cf.* section 1.2.). The temporal-midpoint and middle-third-mean methods (see the top and middle panel, respectively) tend to yield comparable VSA values.

It is instrumental to examine not only the vowel space area, but to focus in more detail on selected vowel spaces. That will allow us to compare the extraction methods in a better way. Figure 2 shows vowel spaces, as extracted by the three methods, for four speakers who manifested some noteworthy tendencies or who may be regarded as representing more speakers with similar patterns. The plots, along with the VSA values, confirm what has been written above, namely that vowel spaces are considerably larger when formants are extracted from the articulatory targets. When we compare vowel spaces in read and spontaneous speech, it is clear that the shifts which underlie the overall reduction of vowel space in spontaneous speech are not identical in the four depicted speakers. It is only in speaker HROK (and similarly also in speaker OLET, not shown in

the figure) that all vowels except the long [iː] are centralized when compared with read speech. Most speakers realized the Czech back vowels – [u uː o oː] – with a higher F2 value, which may, in articulation terms, correspond to centralization and/or weaker or absent lip rounding. However, the close back vowels [u uː] appear to be pronounced in a more peripheral manner in spontaneous speech by speaker JARK, whose vowel space is otherwise extremely small, and also by speakers BRAD, and KALT and NVAT (not shown in Fig. 2). Most speakers also produce more open [a aː] vowels in read speech (in addition to BRAD and HROK in Fig. 2, this applies to another five speakers).

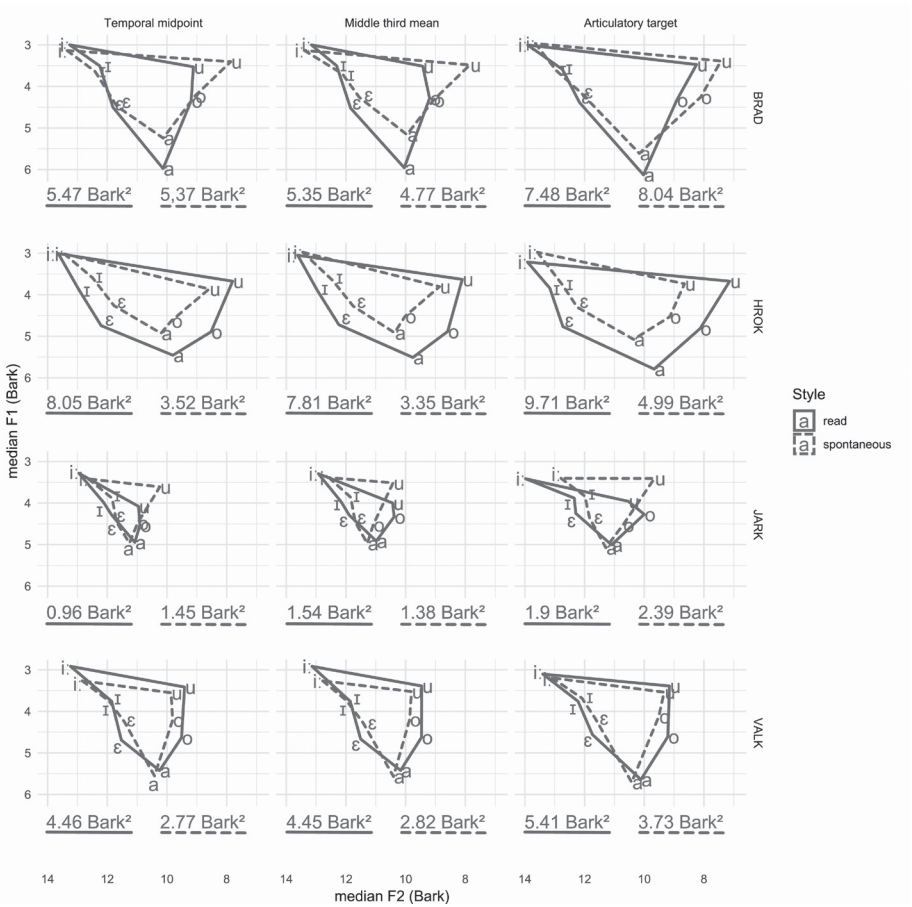

Figure 2. Vowel space of four speakers in read and spontaneous speech, using three methods of formant extraction. The points correspond to medians of F1 and F2. VSA values are provided below each plot.

It is to be expected that median values conceal considerable variability in the data. In Figure 3, we therefore provide another look at the vowel spaces of the four selected speakers in read and spontaneous speech. For the sake of easier comparison, only one extraction method – data based on the middle third mean of each vowel – is shown (as

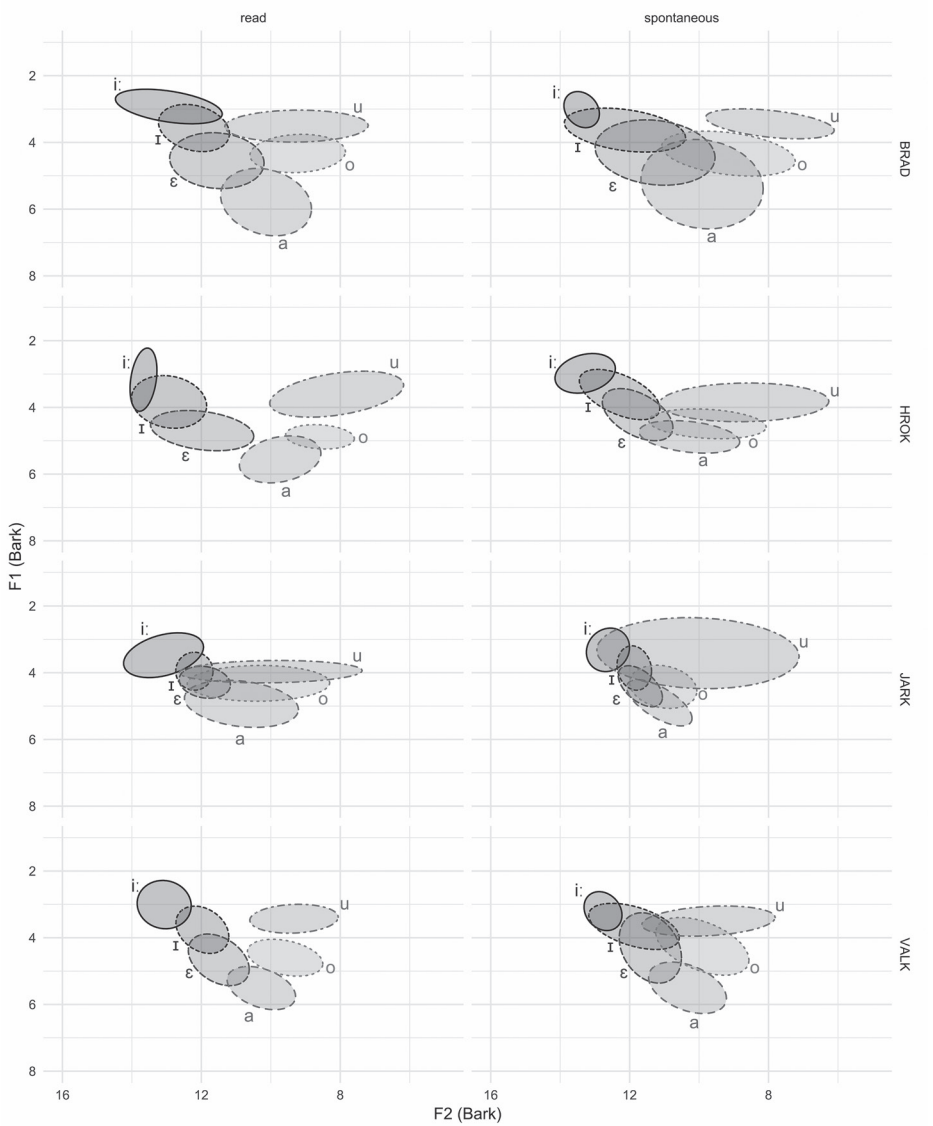

Figure 3. *Vowel spac*e of four speakers in read and spontaneous speech, using the middle third mean method of formant extraction. The ellipses correspond to 68% of the data.

formant extraction from the middle third of a vowel appears to be the most ecologically valid method; see section 1.2).

It is not surprising that there is less overlap between the distributions of formants of individual vowels in read speech. In addition, the overall display of vowel quality distribution in Figure 3 reveals an interesting detail, namely the huge variability of the [u uː] vowels, especially in F2. While the analyses of Skarnitzl and Volín (2012)

indicated a possible change in Common Czech, with the short [u] becoming slightly more centralized than the long [u:], a closer analysis of our data indicates considerable inter- and intra-speaker variability, as shown in Fig. 4. One can see that in most of our speakers the long [u:] is, on average, more peripheral (i.e., has lower F1 and F2 values),

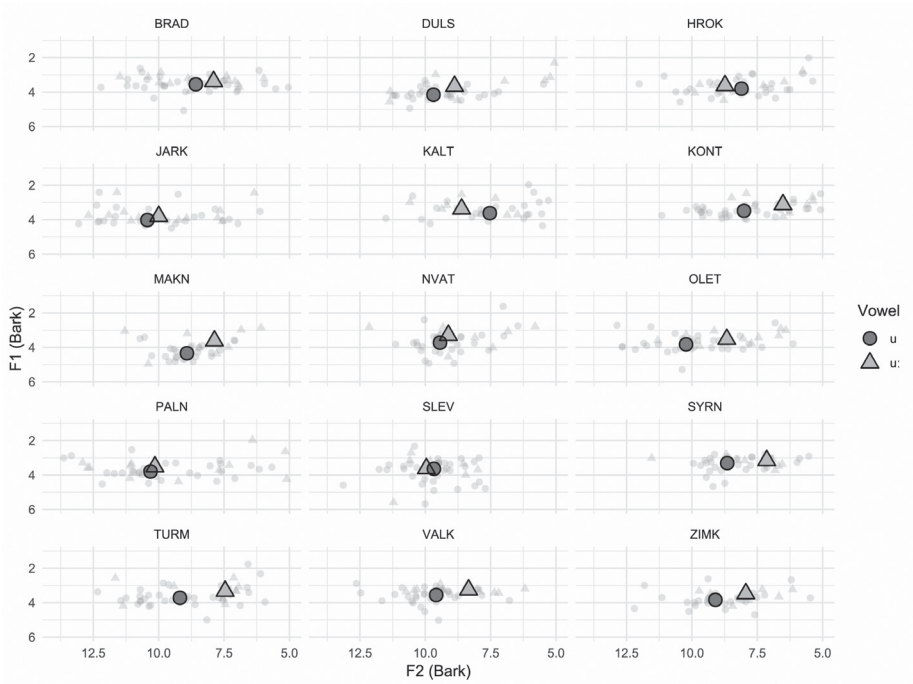

Figure 4. [u] and [u:] F1 and F2 values in individual speakers; the smaller and less opaque points represent single realisations, while the bolder points reflect median values.

but there are several exceptions. Along the horizontal (F2) axis, both vowels manifest considerable variability; indeed, auditory inspection of [u u:] confirmed salient fronting in some tokens.

4. General discussion and conclusion

In our study, we examined variability of vowel space across speakers, in two different speaking styles within a speaker, and using three formant extraction methods. The vowel space characteristics we observed were VSA, i.e., the size of the vowel space area, and distribution of vowel realizations within the acoustic vowel space.

Our results indicate considerable between-speaker differences in vowel space, but also great within-speaker variability between the two speaking styles. We may conclude that vowel space characteristics would not generally be able to differentiate between individual speakers. However, specific speakers may manifest interesting idiosyncratic tendencies which are stable across different conditions and, in comparison with the

comparable population, quite atypical. Speaker JARK in our dataset may serve as an example: as shown in Figure 2, his VSA is markedly smaller compared to other speakers in both speech styles.

Regarding differences between speech styles, speakers' vowel space area tends to be larger in read utterances, reflecting a more distinct, less centralized pronunciation of vowels than in spontaneous speech. Inside the vowel space, there is also an apparent smaller dispersion of values in read speech compared to spontaneous speech, suggesting more consistent articulation of individual vowel qualities (see Figure 3). It must be emphasized that these "divergent" realizations matched perception; in other words, they are not due to faulty formant extraction.

As for the three methods of formant extraction, we calculated formants as the values from the temporal midpoint of a vowel, their mean from the vowel's middle third, and as a single point from the articulatory target defined as the stage where the given formant reached its maximum or minimum. The extraction from the temporal midpoint of vowels represents the most frequent procedure reported in literature (see section 1.2), but it could be argued that a formant value taken from a single time point might frequently correspond to an outlier. This risk can be avoided by taking into account multiple formant values within a vowel, excluding its edges where formants are influenced by the flanking segments, and calculating the mean of those values. However, the two extraction methods did not yield systematically different formant values and VSAs in this study, although the results are certainly not identical (see Figures 1 and 2).

On the other hand, extraction of formants from the vowels' articulatory targets did result in noticeably more extreme values and thus also larger VSAs – at least when considered visually (a quantitative analysis was not an objective of this study). This conclusion is in accordance with Fletcher et al.'s (2015) hypothesis that the articulatory target is not necessarily located in the middle of the vowel's duration. However, it can also be possible that the algorithm set to identify the highest/lowest formant values tends to pick outliers occurring due to faulty formant extraction. Also, the suitability of this extraction method for spontaneous speech can be considered questionable, because – as mentioned above – individuals vowels' pronunciation in reality often corresponds to what appears to be phonetically very different vowel qualities; for example, identifying the articulatory target of an /u/ realization at the F2 minimum can be problematic when the segment's actual pronunciation is closer to a much fronter [y]. This method's validity needs to be further examined; it could be beneficial to identify the placement of identified articulatory targets within vowels and observe whether it is consistent across vowels, and to see if it matches Jacewicz et al.'s (2007) premises (*cf.* section 1.2) or whether its location appears to be random, suggesting the tendency of the algorithm to cling to outliers. In case it shows consistent patterns, it could also be useful to investigate whether the location of the target differs in individual formants, as mentioned by Rose (2015). Moreover, in future research, it could be interesting to focus on LTFs and vowel space density analysis and examine how its outputs correspond to the results of this study.

To conclude, this study has shown that caution should be taken when comparing results of different studies which analyze vowel formants: different extraction methods may provide rather diverging results, and interpretation may thus be less straightforward than it appears.

Acknowledgements

The work was supported by the grant SVV 2020 – 260555 realized at the Charles University, Faculty of Arts, and by the European Regional Development Fund-Project "Creativity and Adaptability as Conditions of the Success of Europe in an Interrelated World" (No. CZ.02.1.01/0.0/0.0/16_019/0000734). We would also like to thank Tomáš Bořil for his assistance with data extraction.

REFERENCES

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

Cavalcanti, J. C., Eriksson, A., & Barbosa, P. A. (2021). Acoustic analysis of vowel formant frequencies in genetically-related and non-genetically related speakers with implications for forensic speaker comparison. *Plos One, 16*(2), e0246645.

Fletcher, A. R., McAuliffe, M. J., Lansford, K. L., & Liss, J. M. (2015). The relationship between speech segment duration and vowel centralization in a group of older speakers. *Journal of the Acoustical Society of America, 138*(4), 2132–2139.

Fletcher, A. R., McAuliffe, M. J., Lansford, K. L., & Liss, J. M. (2017). Assessing vowel centralization in dysarthria: A comparison of methods. *Journal of Speech, Language, and Hearing Research, 60*(2), 341–354.

Fuchs, S. (2017). Changes and challenges in explaining speech variation: A brief review. Available at: https://www.researchgate.net/publication/320991961_Changes_and_challenges_in_explaining_speech_variation_A_brief_review.

Jacewicz, E., Fox, R. A., & Salmons, J. (2007). Vowel space areas across dialects and gender. In *Proceedings of the 16th ICPhS*, 1465–1468.

Machač, P., & Skarnitzl, R. (2009). *Principles of phonetic segmentation.* Epocha.

Nolan, F., & Grigoras, C. (2005). A case for formant analysis in forensic speaker identification. *International Journal of Speech, Language and the Law, 12*(2), 143–173.

Pettinato, M., Tuomainen, O., Granlund, S., & Hazan, V. (2016). Vowel space area in later childhood and adolescence: Effects of age, sex and ease of communication. *Journal of Phonetics, 54*, 1–14.

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

R Core Team (2021). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. Available at: https://www.R-project.org/.

Reetz, H., & Jongman, A. (2009). *Phonetics: Transcription, production, acoustics, and perception*. Blackwell.

Rose, P. (2015). Forensic voice comparison with monophthongal formant trajectories – a likelihood ratio-based discrimination of "schwa" vowel acoustics in a close social group of young Australian females. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4819–4823.

Simpson, A., & Ericsdotter, C. (2007). Sex-specific differences in f0 and vowel space. In *Proceedings of the 16th ICPhS*, 933–936.

Skarnitzl, R., Vaňková, J., & Bořil, T. (2015). Optimizing the extraction of vowel formants. In: Niebuhr, O. & Skarnitzl, R. (Eds.), *Tackling the complexity in speech*, 165–182. Charles University, Faculty of Arts.

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

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

Story, B. H., & Bunton, K. (2017). Vowel space density as an indicator of speech performance. *Journal of the Acoustical Society of America, 141*(5), EL458–EL464.

Šimáčková, Š., Podlipský, V. J., & Chládková, K. (2012). Czech spoken in Bohemia and Moravia. *Journal of the International Phonetic Association, 42*(2), 225–232.

Tykalová, T., Škrabal, D., Bořil, T., Čmejla, R., Volín, J., & Rusz, J. (2021). Effect of Ageing on Acoustic Characteristics of Voice Pitch and Formants in Czech Vowels. *Journal of Voice, 35*(6), 931.e21–931.e33.
Weirich, M., & Simpson, A. (2013). Investigating the relationship between average speaker fundamental frequency and acoustic vowel space size. *Journal of the Acoustical Society of America, 134*(4), 2965–2974.
Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. Available at: https://ggplot2.tidyverse.org/.

Alžběta Houzar
Institute of Phonetics
Faculty of Arts, Charles University
Prague, Czech Republic
E-mail: alzbeta.houzar@ff.cuni.cz

Radek Skarnitzl
Institute of Phonetics
Faculty of Arts, Charles University
Prague, Czech Republic
E-mail: radek.skarnitzl@ff.cuni.cz

2022 ACTA UNIVERSITATIS CAROLINAE PAG. 97–113
PHILOLOGICA 1 / PHONETICA PRAGENSIA

DESCRIPTION OF F0 CONTOURS WITH LEGENDRE POLYNOMIALS

MICHAELA SVATOŠOVÁ, JAN VOLÍN

ABSTRACT

Phonetic research has developed both impressionistic and more objective means of describing the basic units of intonation. The quantification involved in the approaches based on acoustic measurements provides more detail and it is a necessary prerequisite for the comparability and replicability of the results of different studies. In addition to having these characteristics, a proper description of intonation should be comprehensible and meaningful. This article presents a method for describing melodic contours using Legendre polynomials, which yields a few coefficients that capture the basic properties of the analysed contour (e.g. level or slope). This approach thus connects objectivity and quantitative precision with common linguistic concepts. The article also proposes the use of Legendre polynomials for the description of traditionally recognized Czech *melodemes* through the analysis of schemes reported in the literature. Further research on real material could verify the validity of these categories and the usefulness of the method itself.

Key words: intonation, fundamental frequency, Legendre polynomials, polynomial modelling

1. Introduction

Despite differing substantially, various models of intonation share a common goal. They aim to simplify the enormous variability of melodic contours produced by speakers into a limited number of perceptually distinctive categories. This effort involves two tasks – identifying the relevant categories and characterising them appropriately in phonetic terms. The first accounts of intonation were based on careful listening, which is an accessible method that considers perceptually relevant changes in F0. Nevertheless, the impressionistic descriptions formulated as verbal labels (e.g. *fall*, *rise-plateau*) or autosegmental labels (e.g. H*, L+H*) suffer from subjectivity and vagueness.

When instrumental measurements of fundamental frequency became available, new methods emerged that attempted to quantify melodic patterns objectively, e.g. by stating the size of a melodic step in semitones. Experiments have shown that listeners perceive intonation only in the central parts of vowels (Hermes, 2006: 13–15), allowing for one value to represent the pitch of a short syllable. The reduction of a contour into

https://doi.org/10.14712/24646830.2022.31

a set of points connected by lines was licensed by the close-copy stylization approach ('t Hart et al., 1990). On the other hand, the models produced by Fujisaki (1983), Taylor (1994) or Hirst et al. (2000) used complex equations in order to reconstruct F0 contours. Previous studies have also approximated F0 contours with polynomial equations (Andruski & Costello, 2004; Volín & Bořil, 2014). These approaches usually model the original contours more accurately at the expense of interpretability. A compromise is therefore sought that adequately captures the data but remains easily understandable.

This article presents a method for the description of melodic patterns with Legendre polynomials, which provide a quantification that is more linguistically meaningful than conventional polynomial approximations. This approach was already used for the analysis of British nuclear tones (Grabe et al., 2007) and it was applied also to German (de Ruiter, 2011) and Czech (Volín et al., 2017). Furthermore, the research on Czech has exploited Legendre coefficients in the field of automatic speech processing, where they were shown to effectively parameterize the nuclear patterns and improve the prosody of the TTS synthesis (Matura & Jůzová, 2018). Section 2.1 introduces the first four Legendre polynomials (as a subset of the whole Legendre polynomial family) and their coefficients, which relate them to complex curves. The practical steps constituting the process of obtaining the coefficients from the F0 contour are outlined in Section 2.2. The meaning of the coefficients is discussed using real examples in Section 2.3. The following Section 3 then suggests the application of Legendre coefficients in the description of Czech nuclear contours. This demonstration with schematic patterns could serve as a starting point for other studies that could test this approach on real material.

2. Legendre polynomials

2.1 Mathematical basis and modelling of curves

Polynomials are mathematical functions that can be used for describing curves. Legendre polynomials are named after the French mathematician Adrien-Marie Legendre, who discovered them in 1782. They are defined in the interval [−1, 1] and normalized to $L_n(1) = 1$. The first four polynomials are shown in Figure 1 (all figures in this article were created with the R packages *tidyverse*, *grid* and *gridExtra* (Auguie, 2017; R Core Team, 2022; Wickham et al., 2019)). The equation of the Legendre polynomial of the *n*-th degree can be derived from the general formula given in (1). The degree of a polynomial expresses how many changes it can describe. The first polynomial in Figure 1 is constant and it has a degree of 0. The rising line of L_1 already captures one change (from low to high values), which corresponds to the first degree, while the subsequent polynomials of higher degrees turn their directions more times.

$$L_n(x) = 2^n \sum_{k=0}^{n} x^k \binom{n}{k} \binom{\frac{n+k-1}{2}}{n} \tag{1}$$

These polynomials represent basic building blocks for creating more complex curves. Combining pairs of polynomials (by adding together values that correspond to each other on the *x*-axis) yields curves shown in Figure 2. The panel a) illustrates the combination

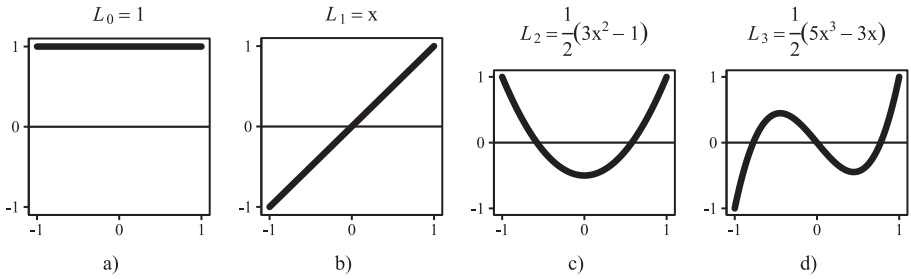

Figure 1. The first four Legendre polynomials (L_0–L_3) with their equations.

of the first two polynomials. While the values of L_1 in its basic form range from −1 to 1 (on the *y*-axis), the addition of L_0 makes them range from 0 to 2, because L_0 has a constant value of 1 in the whole interval. Adding L_0 to any other polynomial would also shift the given polynomial on the *y*-axis, but its shape would remain the same. The curve in the second panel retains the cup-shape of L_2, but it also has a clear rising tendency overall due to the presence of L_1. In the panel c), the direct rise of L_1 is modified by the wave shape of L_3. Finally, the last panel d) shows that summing the nearly opposite values in the first halves of L_2 and L_3 produces values around zero, while their similarly rising shape toward the end results in a more prominent rise.

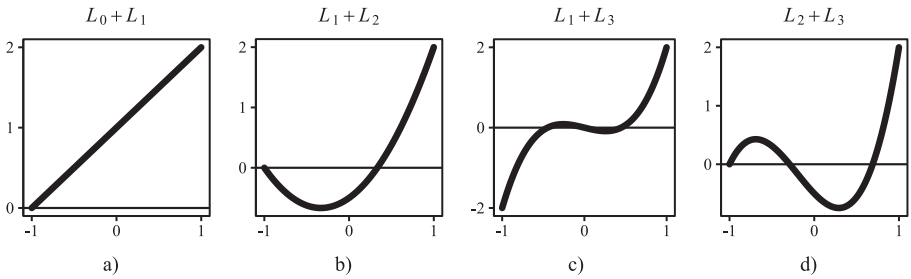

Figure 2. Combinations of pairs of Legendre polynomials.

Each polynomial L_n can be multiplied by the coefficient c_n. The basic polynomials in Figure 1 are not accompanied by any number, which implicitly refers to $c_n = 1$. A different value of c_0 simply makes L_0 represent a different constant, as shown in the panel a) of Figure 3, where $c_0 = 1.8$. The other coefficients affect the span of their respective polynomials. Lowering c_1 to 0.5 produces halved values across the whole interval of L_1, as illustrated in the second panel. Multiplying the polynomials by negative numbers creates curves that are mirror-shape images (according to the *x*-axis) of their counterparts with positive coefficients. Negative values of c_2 therefore lead to dome-shaped curves instead of the cup-shaped ones that were presented so far. This is shown in the panel c), where the coefficient is not only negative, but also of higher absolute value ($c_2 = -2$, compared to $c_2 = 1$ in Figure 1), which is reflected in the wider range of its values. Similarly, a negative c_1 would produce a falling line and a negative c_3 corresponds to a falling-rising-falling shape.

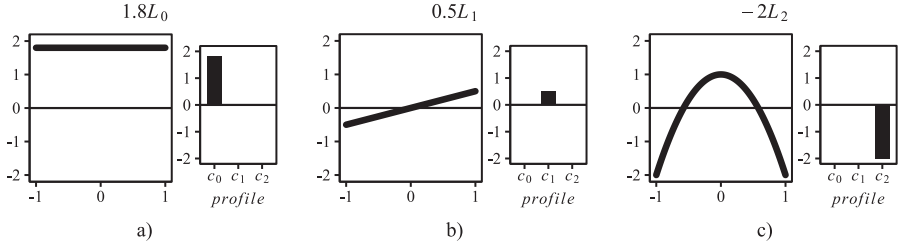

Figure 3. The first three Legendre polynomials multiplied by different coefficients. The *profiles* (on the right side of each panel) show the coefficient values.

The two variations just presented imply the possibility to multiply each basic polynomial with a specific coefficient and add them together. Since the precise values of individual coefficients are no longer easily recognizable from the complex curve, they can be summarized in a *profile* accompanying the curve, as will be done in the rest of the figures in this article. Figure 3 contains three simple *profiles* that graphically depict the respective coefficients. The coefficients of polynomials that are not part of a given curve equal zero. It is therefore sufficient if the *profile* includes coefficients from c_0 up to the last coefficient with a nonzero value. Figure 4 illustrates some of the curves that can be modelled using only two polynomials (L_1 and L_2), but in different ratios. The panel a) starts with a simple fall that corresponds to the polynomial L_1 multiplied by a negative coefficient ($c_1 = -1$). In addition to L_1, the curves in the following panels also include the polynomial L_2 modified by negative values of the coefficient c_2 (these produce dome-shaped curves as in the panel c) of Figure 3). As the relative magnitude of c_2 compared to c_1 gradually increases in panels a) – e), the falling L_1 transforms into the dome-shaped L_2. The curves in panels f) – i) contain a positive c_1, making the overall slope rising. Mirror-shaped images of the first eight curves could be modelled using opposite values of c_1 and c_2 (as indicated in the last four panels), forming a transition from a rise through a cup-shaped parabola to a fall.

Figure 4 shows that curve shapes are determined by the ratios between L_1 and L_2 (and possibly other higher coefficients). In order to compare various curve shapes, relative coefficients (rc_n) can be calculated from the raw ones (c_n) by the formula given in (2), where N stands for the highest degree of a polynomial with a nonzero coefficient. It transforms all coefficients with nonzero values to make the sum of their absolute values equal 1. The first coefficient (c_0) is excluded from this conversion, because it shifts the curve on the y-axis, but it is not related to its shape. The relative coefficients do still express the positive or negative orientation of their respective polynomials. However, they also indicate to what extent do the individual polynomials contribute to the overall shape of the modelled curve. The *profiles* in Figure 4 actually present relative coefficients (the sum of their absolute values equals 1).

(2) $$rc_n = \frac{c_n}{\sum_{i=1}^{N} |c_i|}$$

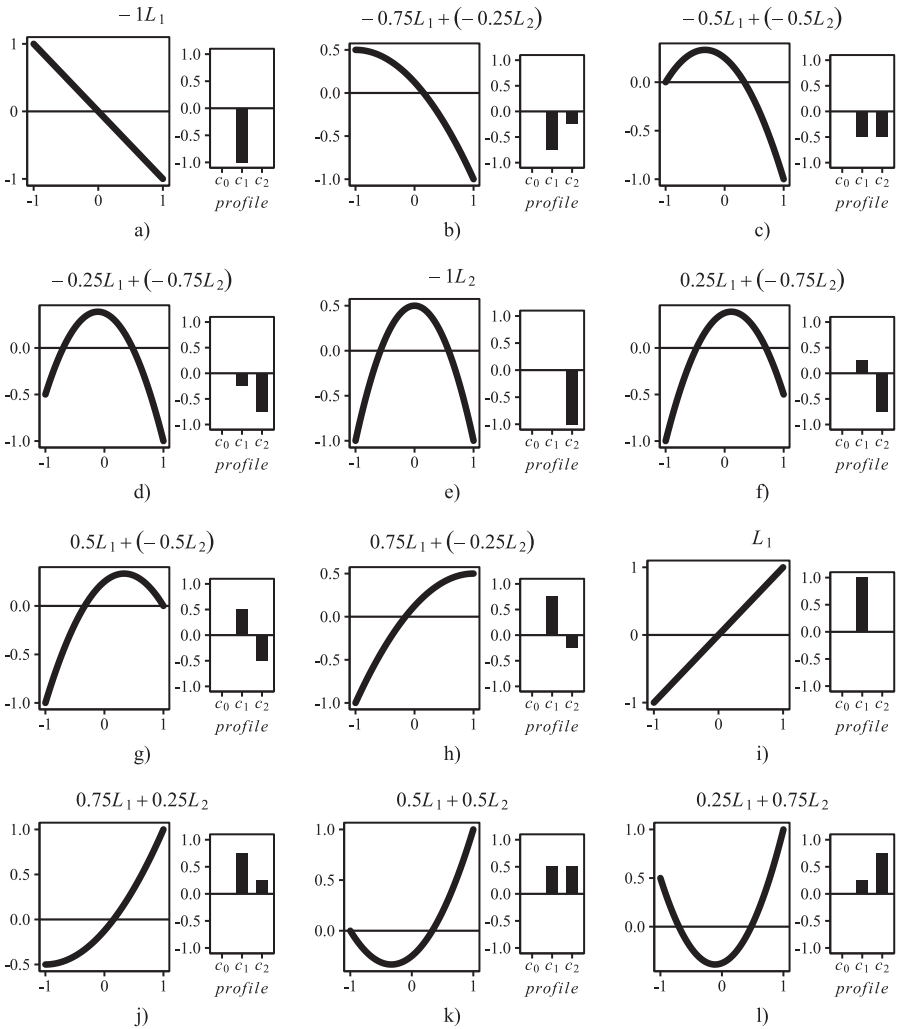

Figure 4. Curve shapes produced by combinations of the second (L_1) and third (L_2) Legendre polynomial. Each polynomial is multiplied by a different coefficient (their values are shown in the *profiles* on the right side of each panel).

An important feature of Legendre polynomials is their orthogonality. In mathematical terms it means that the inner product of each two polynomials equals zero. It refers to the fact that each polynomial captures a property that is not explainable by any other polynomial. The average value of a given curve can be described only using L_0, since all the other polynomials have an average of zero. Similarly, the linear slope is only reflected in L_1, because the linear regression of the other polynomials equals zero. The polynomials of higher degrees describe unique characteristics in the same manner. Thanks to orthogonality, any curve in the [−1, 1] interval can be formed by summing polynomials

of different degrees, each multiplied by a specific coefficient, although the modelling of more complex curves (containing numerous or abrupt changes) requires using polynomials of higher degrees. From the opposite perspective, it is possible to decompose any curve (e.g. an interpolated F0 contour) into a set of Legendre polynomials multiplied by different coefficients (having different "amplitudes"). This is analogical to Fourier analysis of a sound wave, which works with cosine functions of various frequencies instead of Legendre polynomials. Identifying the values of these coefficients lies at the core of the analysis presented in Section 2.2. Since the inner product of any pair of Legendre polynomials equals zero, the coefficient of each polynomial can be calculated from the inner product of that polynomial and the analysed curve. Another consequence of orthogonality is the independence of the coefficients, which means they are not correlated and can be statistically evaluated as separate variables within one analysis.

The first few polynomials are sufficient for the analysis of intonation, because they capture the main melodic movements and ignore microprosodic effects. In other words, they model the relatively simple underlying shape of the F0 contour. Specifically, in nuclear patterns that usually span over a few syllables, a higher number of changes is not expected. For convenience, the first four coefficients are referred to as AVERAGE (c_0), SLOPE (c_1), PARABOLA (c_2) and WAVE (c_3) following Grabe et al. (2007). Their further advantage over other methods of polynomial modelling (e.g. least squares approximation) is that they can be interpreted in linguistic terms. As mentioned earlier, the AVERAGE has a special position, because it expresses the mean value of the curve, while all the other coefficients are related to its shape. The meaning of AVERAGE and units in which it is expressed depend on the method of normalization of the original F0 contour (described in detail in Section 2.2), but it is connected to the position of the nuclear pattern in the pitch range. The relative ratios of SLOPE, PARABOLA and WAVE affect the shape of the curve (as illustrated in Figure 4), while the absolute values of these coefficients reflect its span (compressed or expanded). A feature that is not inherently captured by this method is the temporal dimension, since each contour needs to be transformed to the interval [−1, 1].

2.2 Analysis of F0 contours

This section describes the necessary steps that have to be undertaken in order to obtain the Legendre coefficients of a given melodic contour. Some of them are common in intonation research, while others are required specifically by this type of analysis.

First of all, the analysis domain has to be chosen. Stress-groups usually consist of a few syllables bearing relatively simple melodic movements. These can be adequately captured by a few coefficients and therefore seem as a reasonable choice. The decision about the appropriate domain should be guided by the research question and by the findings from the previous research on the given language. For example, it might be useful to include the pre-stressed syllable, if its relative pitch plays a role in distinguishing various patterns, as was done for German in de Ruiter (2011). However, longer units such as prosodic phrases could be modelled as well, if the differences in interpretation are taken into account. One caveat arising from this approach might be the diversity of attested patterns. The range of possibly distinct contours expands with every additional

syllable of the analysed domain, which also affects the amount of data necessary for their adequate description.

The present analysis is based on the measurement of F0 alone. It differs in this respect from the method utilized by Grabe et al. (2007), which additionally required the parameters of intensity and periodicity. These were used for weighting the importance of particular parts of the F0 contour. The idea underlying their approach derives from the finding that listeners do not pay equal attention to variations of pitch in different segments. More sonorant phones like vowels appear to form the basis of the perceived melodic patterns, while the pitch in voiced obstruents is ignored with regard to intonation (Hermes, 2006: 13–15). The method described here proposes an alternative way of reflecting this knowledge through the exclusion of all F0 values in the irrelevant regions from the analysis. Although it provides only a categorical distinction (values are either used or deleted) instead of a gradual scale, this approach avoids further errors that are related to the measurements of other parameters.

Since listeners' sensitivity to pitch variations seems to be language-specific, the choice of particular parts of the F0 contour for the analysis should be theoretically grounded. Generally, the F0 values would be retained in vowels and discarded in consonants, although some languages might exploit also the regions occupied by sonorants. This elementary distinction could be further refined to eliminate some of the microprosodic effects. This means extracting only certain parts of each vowel, defined either absolutely (e.g. starting 10 ms after the beginning of the vowel) or relatively (e.g. using the middle third of its duration). Comparing both approaches might show if the simpler method is robust enough or whether the further adjustments need to be made. It is obvious that in any case, the F0 contour should be annotated at the level of segments.

From the practical point of view, the F0 contour in the domain of interest has to be extracted and corrected for errors like octave jumps and missing values, which is a standard procedure for studies concerning pitch. The values calculated in Hertz are then commonly converted to semitones (ST), which applies also for the present analysis, because this unit is perceptually more relevant. Semitones are used (instead of octave ratios as in Grabe et al., 2007) for two main reasons – their values are easier to interpret and they are well known. Nevertheless, the two units are mutually convertible (for comparative reasons) by simply dividing the values in semitones by 12 and vice versa, which holds for the coefficients as well.

Furthermore, the contours can be normalized to allow comparisons between speakers and utterances. The interpretation of AVERAGE (the first coefficient) follows directly from the chosen reference. Subtracting the mean F0 of each speaker implies that the coefficient is to be understood in relation to it. For example, if the contours of a speaker are expressed in semitones with the reference of 100 Hz (his mean F0), then the AVERAGE of 1.5 corresponds to 109 Hz, which is 1.5 ST above 100 Hz. Depending on the research question, an alternative reference for the normalization might be chosen (e.g. the mean F0 of the given utterance).

Finally, the coefficients are calculated using the method implemented in the *rPraat* package for the R software (Bořil & Skarnitzl, 2016; R Core Team, 2022). As already mentioned, the procedure differs from the one described in Grabe et al. (2007), but it derives from the orthogonality of Legendre polynomials. The algorithm takes the adapted F0

contour (normalized and including only the relevant values) and performs the following operations on it. First, the contour is linearly interpolated into 1000 points to ensure equivalent sensitivity in the whole analysed domain. Secondly, the time scale is transformed into the interval [−1, 1] in which Legendre polynomials are defined. This means that the coefficients alone are unable to capture the stretching or compressing of an identically shaped contour in the temporal dimension. Lastly, the coefficients are calculated from the inner products of the transformed contour and the respective polynomials.

The outlined method is summarized in these four steps:

1. selection of the analysis domain and its relevant parts
2. extraction of the F0 contour (in semitones)
3. normalization
4. calculation of the coefficients

An example of a practical application of this procedure is provided here, using a short polarity question [ˈɻɛknɛtɛ jɪm ˈʧo sɪ ˈmɪsliːtɛ] ("Will you tell them what you think?") produced by six speakers. The nuclear contour spanning the last stress-group [ˈmɪsliːtɛ] was chosen as the analysis domain, limiting the relevant parts only to vowels, which form the nuclei of the three syllables. F0 estimates in 10 ms intervals were extracted in Praat (Boersma & Weenink, 2022) using the autocorrelation method with standard settings and then manually corrected for octave jumps and missed voicing regions. All values in Hertz were converted to semitones with the reference of 1 Hz. The speakers' means (obtained from a collection of their utterances) were then subtracted from the respective contours.

Figure 5 shows all six nuclear contours. The black points represent the extracted F0 values, while the interpolated values (also used for the analysis) are coloured in grey. The Legendre coefficients (c_0–c_3 in this case) are presented in the *profiles* on the right side of each panel. The black curves are models based solely on these four coefficients. Their values are summarised in Table 1 together with their relative counterparts and explained in the next section.

2.3 Interpreting the coefficients

So far, only coefficients relating to isolated curves were discussed. However, working with real data usually requires comparing multiple contours. This section therefore provides a more detailed description of the relationship between Legendre coefficients and the curves they represent. It also explains how similarities of contour shapes translate into the coefficient values.

The speaker S1 realized the analysed nuclear pattern as a *rise* (with the main melodic step located between the first and the second syllable), as shown in the top left panel of Figure 5. The accompanying *profile* reveals that its most prominent coefficient is the positive SLOPE, followed by the negative PARABOLA, which reflect the rising and dome-shaped appearance (see Figure 3 above for the individual polynomials and Figure 4 for their combination). The minor AVERAGE indicates that the whole contour is located 0.1 ST below the speaker's mean pitch.

The coefficients of S1's contour can be compared to those of another speaker, S2. Turning to both top panels of Figure 5, it can be seen that the extracted F0 values

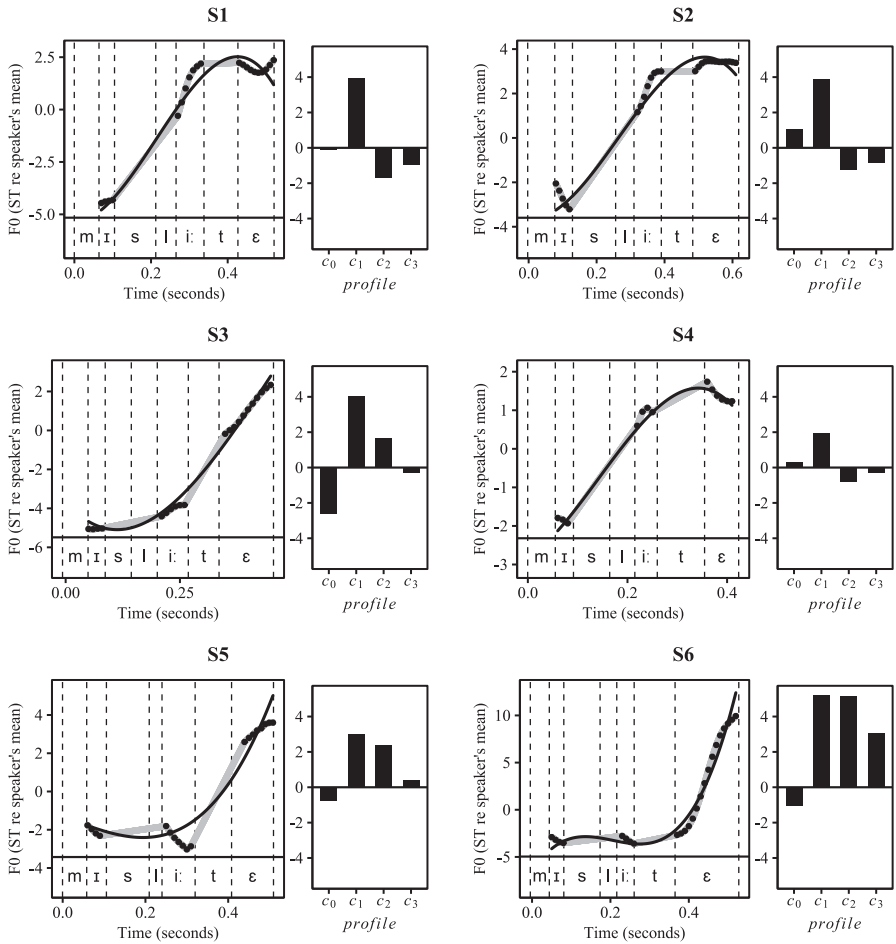

Figure 5. The nuclear contours on the stress-group [ˈmɪsliːtɛ] produced by six speakers. Each panel includes the extracted F0 values (black points), the interpolated values (in grey) and a curve constructed from the first four Legendre coefficients (these are shown in the *profiles* on the right side of each panel).

resemble each other a lot. Conveniently, the similarity is preserved in the coefficients of these contours. Excluding the AVERAGE that is not related to the shape, but rather signals the position of the nuclear pattern in the pitch range (1.1 ST above S2's mean pitch), the coefficients do not differ from each other by more than 0.4. In contrast, S3 produced a pattern that could be called a *late rise*, realizing the main melodic step between the second and the third syllable. It results in a considerable change in PARABOLA, which switches to a positive value. It captures the cup-shaped pattern that is present in the contour, although reduced by the more prominent SLOPE. It also becomes clear that the coefficients should not be interpreted in isolation. The similar SLOPE of the first three contours does not imply their resemblance in the overall

shape. Nevertheless, it still holds that they all include overall rising, which is exactly what SLOPE represents. However, the type of rising is specified only in combination with higher coefficients.

The contour of S3 also features a lower AVERAGE. While this coefficient is autonomous in the sense that it merely distinguishes between realizations of an identical shape on different levels in the pitch range, it also interacts with the shape in one respect. The first and third contour begin and end with comparable F0 values of approximately −4.5 and 2 ST, yet their values of AVERAGE differ by 2.5 ST. The reason is the shape of the contour (here specifically the position of the second syllable), because all interpolated points in the analysed interval contribute equally to the value of AVERAGE. The AVERAGE will always be lower for contours with a low middle part than for those with a high one, even though they have the same values on the edges. It remains to be tested whether it reflects the fact that listeners perceive more low or high values in the whole contour. However, this effect should be taken into account in the interpretation.

Table 1. The Legendre coefficients (raw on the left, relative in italics on the right) of the six nuclear contours from Figure 5.

	AVERAGE	SLOPE	PARABOLA	WAVE	*SLOPE*	*PARABOLA*	*WAVE*
speaker	c_0	c_1	c_2	c_3	rc_1	rc_2	rc_3
S1	−0.1	3.9	−1.7	−0.9	0.60	−0.26	−0.14
S2	1.1	3.9	−1.3	−0.8	0.65	−0.21	−0.14
S3	−2.6	4.0	1.7	−0.3	0.67	0.28	−0.05
S4	0.3	1.9	−0.8	−0.3	0.63	−0.27	−0.10
S5	−0.8	3.0	2.4	0.4	0.52	0.41	0.07
S6	−1.0	5.2	5.2	3.1	0.39	0.38	0.23

Despite the fact that the contour produced by S4 follows the same pattern as the first two, its coefficients (especially SLOPE) substantially differ. This is due to the narrower span it covers, because the *rise* stretches only around 3.5 ST compared to approximately 6.5 ST in the previous contours. The relative coefficients in Table 1 clarify the similarity of S4's contour to those of S1 and S2. It lies in the approximately 63% share of positive SLOPE and 23% share of negative PARABOLA. The inclination of S3's contour to the shape of a *late rise* is therefore mainly caused by the positive PARABOLA, although it is present in a similar ratio. The relative coefficients also show that SLOPE forms the most important component in these contours, leaving only half as much space to PARABOLA. Notice that the ratios between coefficients can be to a certain extent visually assessed from the *profiles* alone, even if their *y*-axes have the same range and the raw coefficients differ in magnitude.

S5 realised a contour with a similar appearance to the one produced by S3 (a *late rise*), which is reflected in the comparable relative coefficients of these two contours. The

prominence of the SLOPE is reduced in S5's contour (although it is still the strongest component), while the PARABOLA has a greater share than is S3's pattern. This change in the ratios of both coefficients relates to the position of the second syllable. The third panel resembles a straight rising line, while the fifth is more bent (compare with panels j) and k) in Figure 4). An important feature for the differentiation of curve shapes is the polarity of the most prominent coefficients, which applies for SLOPE and PARABOLA here. The contrast of positive and negative WAVE does not affect the shapes dramatically, because its relative values are close to zero.

The last panel of Figure 5 shows another *late rise*. However, the speaker S6 did not produce the melodic step between the second and third syllable (as did S3 and S5), but compressed this movement into the final syllable, which starts at a low pitch and ends high. The modelling of this abrupt change requires the presence of WAVE. Its relative value is three times higher for S6 than for S5 and at the same time the highest of all the contours in Figure 5. Besides allowing for the steep rise, the WAVE also makes the first half relatively flat. Without it, the combination of SLOPE and PARABOLA would result in a fall in that part of the curve (as in the panel k) in Figure 4). In fact, any curve with a steeper or sharper shape than those in Figure 4 necessarily includes WAVE (or other higher coefficients) in a non-negligible ratio. The same tendency can be observed even in the contours of S1 and S2. Their values of relative SLOPE and PARABOLA lie somewhere between those from panels g) and h) in Figure 4, but they rise in a straighter manner due to 14% of negative WAVE.

3. Modelling Czech nuclear patterns with Legendre polynomials

3.1 Connecting Legendre polynomials to linguistic categories

Previous sections have described the method which allows for the simplification of a F0 contour into a few Legendre coefficients that capture its basic properties. Nevertheless, these individual numbers do not represent the ultimate goal of phonetic research. Returning to the introduction, the aim of the analysis is to describe intonation patterns generally, which involves classification into various categories. Each category includes a range of possible realizations, while remaining distinct from other categories in various ways. The difference between declarative and interrogative utterances is a commonly mentioned one, but contours of different types can among other things also signal the speaker's dialect, thus expressing the indexical function.

The most thoroughly studied unit within the intonation of Czech is the nuclear pattern (*melodeme*), which is considered the most information-laden. Three functionally distinct categories are distinguished – conclusive, interrogative and continuative patterns (Daneš, 1957: 38-54; Palková, 1994: 307–315), each consisting of several contour types (*cadences*). Their traditional descriptions were based on auditory assessment, resulting in schematic stylizations using four pitch levels (Daneš, 1957: 53–54). It would be desirable to experimentally test their perceptual validity. However, the controlled design of

test items requires quantitative characteristics of these types. Legendre coefficients might serve as a convenient tool in this respect. A given category can be described with a model contour created from the average coefficients of the contours belonging to that category. For illustration purposes, the following sections present an experiment indicating approximate coefficient values which could be expected for some of the traditionally reported contour types.

3.2 Method

The three categories of *melodemes* were represented by schematic patterns taken from Palková (1994: 309–315) and spanned over a three-syllabic stress-group, which allows for a greater variability of contour shapes. The patterns are summarised in Table 2. In order to analyse them with Legendre polynomials, F0 contours based on these schemes were created in Praat. The four levels were set to 305, 265, 230 and 200 Hz, which yielded equidistant pitch levels after the conversion to semitones (approximately 2.4 ST apart). The values of the highest and lowest level were chosen to reflect a possible human pitch range, which produces coefficient values comparable in magnitude to those that could be obtained from real recordings. The target values were placed in the centres of vowels in a simple CVCVCV template (assuming the same duration for all segments) and then interpolated quadratically with the built-in function in Praat. The edge values in the first and last vowel were adjusted to produce a mean F0 (in these vowels) equivalent to the desired levels. The analysis domain thus corresponded to a stress-group and the relevant parts used for the analysis were limited to the vowels.

Table 2. Schemes of Czech nuclear patterns based on Palková (1994: 309–315). Number 1 represents the highest pitch level, number 4 the lowest; * marks the stressed nuclear syllable and the number in brackets denotes the level of the pre-nuclear syllable.

	conclusive patterns (CCL)	interrogative patterns (INT)	continuative patterns (CNT)
1	(1) 2* 3 4	(2) 4* 4 2	(3) 4* 3 2
2	(2) 3* 2 4	(2) 4* 1 2	(4) 2* 1 2
3		(4) 1* 1 2	(4) 1* 1 3

The pitch of the pre-nuclear syllable was chosen as the reference value for normalization for two reasons. First of all, the mean F0 of a speaker or utterance could not be used, since the contours were created artificially in isolation. Secondly, the relative position of the stressed syllable and the preceding (pre-nuclear) syllable is argued to differentiate various patterns and therefore represents a relevant component of the whole contour (Daneš, 1957: 51). Legendre coefficients were calculated in *rPraat* with the method explained in detail at the end of Section 2.2. Their relative counterparts were obtained using the formula presented in Section 2.1.

3.3 Interpretation

Figure 6 illustrates all analysed contours. Similarly to Figure 5, it contains the original F0 values (black points) and the whole interpolated contours (grey lines). The black curves represent the models based on the first four Legendre coefficients, which are shown in the profiles on the right. Both raw and relative coefficients are summarised in Table 3. However, the specific values should not be taken literally, since a few arbitrary decisions had to be made when transforming the schematic representations into analysable F0 contours.

The first two contours are conclusive and they both have a negative SLOPE. It indicates an overall fall, which is a typical property of this category. However, they differ in some respects. The fall in CCL-2 is milder (it has a smaller absolute SLOPE) and it is complemented by a rising-falling element, which is captured by the negative PARABOLA. As can be seen from the relative coefficients, the parabolic shape in fact contributes to the whole contour to a greater extent than SLOPE. On the other hand, the first two interrogative contours are rising, since they both contain a positive SLOPE, although it is not the only polynomial present in them. Nevertheless, considering just the first four contours, the negative or positive SLOPE seems to differentiate between the conclusive and interrogative types.

Table 3. The Legendre coefficients (raw on the left, relative in italics on the right) of the contours from Figure 6. The sum of the absolute values of relative coefficients does not equal 1 in some rows due to rounding.

	AVERAGE	SLOPE	PARABOLA	WAVE	*SLOPE*	*PARABOLA*	*WAVE*
contour	c_0	c_1	c_2	c_3	rc_1	rc_2	rc_3
CCL-1	–4.9	–3.0	0.0	0.2	–0.94	0.00	0.06
CCL-2	–2.2	–1.5	–2.9	0.1	–0.33	–0.65	0.02
INT-1	–3.4	3.0	1.9	–0.2	0.59	0.37	–0.04
INT-2	–0.5	3.0	–3.9	–0.2	0.43	–0.55	–0.03
INT-3	6.6	–1.5	–0.9	0.1	–0.60	–0.37	0.04
CNT-1	0.0	3.0	0.0	–0.2	0.94	0.00	–0.06
CNT-2	5.8	0.0	–1.9	0.0	0.00	–1.00	0.00
CNT-3	5.8	–3.0	–1.9	0.2	–0.59	–0.37	0.04

The advantage of Legendre coefficients over verbal labels manifests itself in the comparison of CCL-2 with INT-2 and CNT-2. They could all be called *rise-falls* based on the relative positions of the three syllables, despite the fact that they are visually and perceptually distinct. A longer specification is then required to capture the different magnitudes of the melodic steps between syllables. The descriptions become much more concise when translated into Legendre coefficients. The most prominent element is the negative PARABOLA, corresponding to the rising-falling skeleton shared by all three

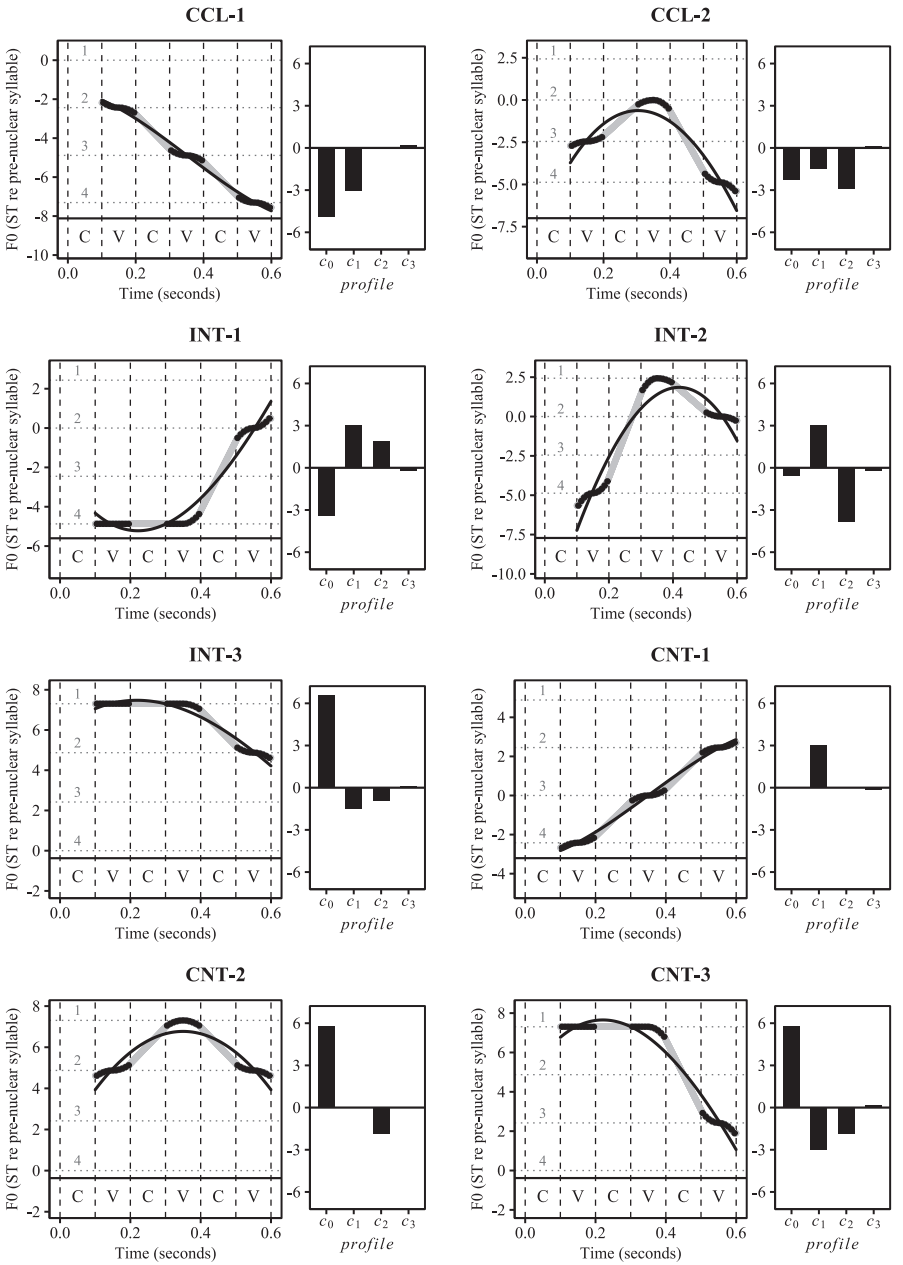

Figure 6. The modelled contours based on the schematic patterns from Table 2. Each panel includes the F0 values (black points), the interpolated values (in grey) and a curve constructed from the first four Legendre coefficients (these are shown in the *profiles* on the right side of each panel). The horizontal dotted lines indicate the four pitch levels.

contours. It is the only constitutive component of the model in the case of CNT-2, but it is complemented by SLOPE in the other two contours. These are distinguished by the values of SLOPE – CCL-2 has a falling tendency, while INT-2 is rising. The SLOPE is thus indirectly signalling the ratio between the two melodic steps.

The first interrogative contour resembles the second one in the rising aspect, but it has a positive value of PARABOLA, which on its own means a fall-rise. However, the relative coefficient shows that it amounts to approximately one third of the whole contour, while the SLOPE has a greater share. This combination lies halfway between the shapes j) and k) in Figure 4 and leads to a plateau (rather than a fall) between the first two syllables followed by a rise. The third interrogative contour diverges from the general pattern, since it is falling, although not prominently in absolute terms (small absolute value of SLOPE). This opposite tendency is compensated by a high AVERAGE that strongly contrasts with the negative or zero values present in the contours discussed so far. The specific values are arbitrary, because they result from the F0 levels chosen during the modelling of the contours, but the ratios between them hold true. The interpretation of AVERAGE depends on the current reference, which is the F0 level of the pre-nuclear syllable here. While the first four contours are located at about the same pitch or below the previous syllable, INT-3 lies higher. This might serve a similar function as the overall rising, since both strategies end at a relatively high pitch.

Interestingly, the same pattern is present in the continuative contours. The first one has a positive SLOPE, which is also dominant relatively. On the contrary, CNT-2 and CNT-3 contain a high AVERAGE, although their SLOPE is zero or even negative. In fact, CNT-3 closely mirrors the relative coefficients of INT-3. Figure 6 shows that the two shapes are alike, but the two contours differ in the level of the last syllable (see Table 2). In other words, CNT-3 covers a wider span. This difference is normalised in the relative coefficients, but retained in the raw coefficients, which are halved for INT-3. The level and span of nuclear patterns might play an important role for the listeners when distinguishing the categories. Although INT-3 and CNT-3 seem to differ only in the span, the present analysis is strongly limited by the four-level schematization and a proper description would require real data.

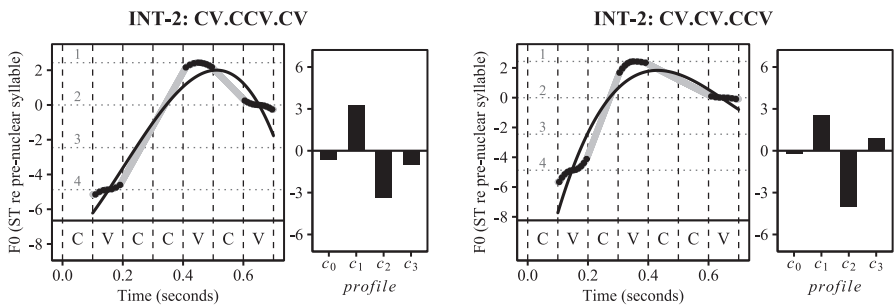

Figure 7. The modelled contours of the INT-2 pattern as combined with two different syllabic templates. Each panel includes the F0 values (black points), the interpolated values (in grey) and a curve constructed from the first four Legendre coefficients (these are shown in the *profiles* on the right side of each panel). The horizontal dotted lines indicate the four pitch levels.

Finally, it can be seen at first glance that WAVE is only marginally involved in all contours, probably due to the regular temporal distribution of the three melodic targets. For comparison, Figure 7 simulates the presence of a consonant cluster before or after the second vowel in contour INT-2. The relative position of the peak is therefore shifted towards the beginning or end of the contour. It leads to four times higher relative values of WAVE (−0.13 and 0.12, respectively) compared to the rc_3 of the original INT-2 contour. These account for the steeper falls at the edges of the modified contours.

4. Conclusion

The article presented a method for the description of F0 contours using Legendre polynomials. The main melodic movements are converted into a few (usually four) coefficients that are linguistically interpretable, while remaining quantitatively precise. They therefore combine the advantages of simple verbal labels and complex mathematical equations. The coefficients capture the elementary properties of the analysed contours and ignore microprosodic effects. Both dimensions of the pitch range are referred to in this approach, since the AVERAGE (the first Legendre coefficient, c_0) relates to the level and the absolute values of the other coefficients reflect the span. Different contour shapes can be easily compared using the relative coefficients, which inherently express the internal temporal distribution of the pitch targets in the analysed contour. However, the total duration of the analysed unit is not accounted for due to the normalization that is required for the calculations. Relating the coefficients to speech tempo thus remains one of the questions for further research.

Section 3 suggested the application of Legendre coefficients in the description of Czech nuclear patterns. However, it only outlined the procedure that should be repeated with natural material in order to explore the differences between the nuclear pattern categories and also the specifics of their subtypes. These studies could compare their results with those observed here for the traditional schemes and test the usefulness of Legendre coefficients in intonology. The following step could turn to the listener and examine the distinctiveness of contour subtypes in perception experiments. A potential systematic relationship between the perceived perceptual differences of F0 contours and the values of their Legendre coefficients would provide further evidence for the relevance of this method.

Acknowledgements

This work was supported by Czech Science Foundation (GAČR), project GA21-14758S.

REFERENCES

Andruski, J. E., & Costello, J. (2004). Using polynomial equations to model pitch contour shape in lexical tones: An example from Green Mong. *Journal of the International Phonetic Association*, *34*(2), 125–140.

Auguie, B. (2017). *gridExtra: Miscellaneous Functions for 'Grid' Graphics*. https://CRAN.R-project.org/package=gridExtra

Boersma, P., & Weenink, D. (2022). *Praat: Doing phonetics by computer* (6.2.07). https://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* (pp. 367–374). Springer International Publishing.
Daneš, F. (1957). *Intonace a věta ve spisovné češtině*. Nakladatelství ČSAV.
de Ruiter, L. E. (2011). Polynomial Modeling of Child and Adult Intonation in German Spontaneous Speech. *Language & Speech*, *54*(2), 199–223.
Fujisaki, H. (1983). Dynamic Characteristics of Voice Fundamental Frequency in Speech and Singing. In P. F. MacNeilage (Ed.), *The Production of Speech* (pp. 39–55). Springer.
Grabe, E., Kochanski, G., & Coleman, J. (2007). Connecting Intonation Labels to Mathematical Descriptions of Fundamental Frequency. *Language & Speech*, *50*(3), 281–310.
Hermes, D. J. (2006). Stylization of Pitch Contours. In S. Sudhoff, D. Lenertova, R. Meyer, S. Pappert, P. Augurzky, I. Mleinek, N. Richter, & J. Schließer (Eds.), *Methods in Empirical Prosody Research*. DE GRUYTER.
Hirst, D., Di Cristo, A., & Espesser, R. (2000). Levels of Representation and Levels of Analysis for the Description of Intonation Systems. In M. Horne (Ed.), *Prosody: Theory and Experiment* (Vol. 14, pp. 51–87). Springer Netherlands.
Matura, M., & Jůzová, M. (2018). Correction of Formal Prosodic Structures in Czech Corpora Using Legendre Polynomials. In A. Karpov, O. Jokisch, & R. Potapova (Eds.), *Speech and Computer* (pp. 387–397). Springer International Publishing.
Palková, Z. (1994). *Fonetika a fonologie češtiny: S obecným úvodem do problematiky oboru* (1. vydání). Karolinum.
R Core Team. (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/
't Hart, J., Collier, R., & Cohen, A. (1990). *A Perceptual Study of Intonation: An Experimental-Phonetic Approach to Speech Melody*. Cambridge University Press.
Taylor, P. (1994). The rise/fall/connection model of intonation. *Speech Communication*, *15*(1–2), 169–186.
Volín, J., & Bořil, T. (2014). General and speaker-specific properties of F0 contours in short utterances. *Acta Universitatis Carolinae – Philologica*, 9–19.
Volín, J., Tykalová, T., & Bořil, T. (2017). Stability of Prosodic Characteristics Across Age and Gender Groups. *Interspeech 2017*, 3902–3906.
Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686–1691.

Michaela Svatošová
Institute of Phonetics
Faculty of Arts, Charles University
Prague, Czech Republic
E-mail: michaela.svatosova@atarien.com

Jan Volín
Institute of Phonetics
Faculty of Arts, Charles University
Prague, Czech Republic
E-mail: jan.volin@ff.cuni.cz

2022 ACTA UNIVERSITATIS CAROLINAE PAG. 115–129
PHILOLOGICA 1 / PHONETICA PRAGENSIA

ONLINE GUIDED PRONUNCIATION PRACTICE HELPS ADULT EFL LEARNERS IMPROVE L2 PROSODY

ŠÁRKA ŠIMÁČKOVÁ, VÁCLAV JONÁŠ PODLIPSKÝ
Palacký University Olomouc

ABSTRACT

This study tests the efficacy of a pronunciation course in developing advanced EFL learners' expressive reading during a semester of online instruction. The course, designed for future English-language professionals, emphasises primacy of perception before production, the importance of noticing phonetic detail, expert and peer feedback, and context-situated tasks. The magnitude of pitch movements and reading tempo were assessed before and after the course for a trained group, who received the pronunciation practice, and a comparison group attending a course about the theory and research of foreign accents in English. Only the Trained group's expressive prosody improved: the learners slowed down their delivery and produced the utterances with a wider pitch range. The results suggest that adult foreign language learners can benefit from pronunciation training in a distance learning environment.

Key words: distance learning, English pronunciation, expressive prosody, pitch range, reading tempo

1. Introduction

1.1 Research background

Prosody is essential to organising connected speech and it plays an important role in effective communication (Hirschberg, 2002). The prosodic structure of an utterance reflects the speaker's organisation of thought (in prosodic phrasing and prominence), the degree of certainty with which they are speaking (in melodic patterns), as well as pragmatic meanings that the speaker may want to convey beyond the lexical meaning of spoken words (such as doubt, surprise or irony). To the listener, prosodic cues indicate what to pay attention to when processing speech, what to anticipate in the upcoming discourse, or when to take their turn in conversation.

Mastering the prosody of a second language (L2) is not an easy feat. L2 prosodic learning is affected by first-language (L1) transfer just like the learning of L2 segments (Mennen & de Leeuw, 2014) and prosodic features of L2 speech often serve as markers of foreignness to a native listener's ear (De Mareüil & Vieru-Dimulescu, 2006). Non-na-

https://doi.org/10.14712/24646830.2022.32

tive prosody may negatively impact the reception of an L2 speaker's message, affecting listeners' interest in what is being said, or reducing comprehension (Kang, 2008; Kang, Rubin, & Pickering 2010). Although research on L2 prosody training is relatively limited, existing studies have shown that foreign language prosody can be improved through training (see Lengeris, 2012 for a review). In the speech of Czech learners of English, the learner population trained in this study, prosodic features (F0 variation and articulatory rate) have been shown to predict accent ratings (Volín and Skarnitzl, 2010), although the narrow pitch range often taken to be typical of Czech-accented English cannot be attributed purely to L1 interference (Volín, Poesová & Weingartová, 2013).

The current study tests the ability of young adult non-immersion Czech EFL learners to change prosodic aspects of their English speech as a result of guided pronunciation practice. Our learners are students at Palacký University training to become English language professionals, and as such are highly proficient and active EFL users. They opted for a 13-week-long pronunciation course and thus can be regarded as motivated to improve their spoken English. Since 2 weeks into the course the first covid lockdown closed all classrooms, the training moved online. It relied primarily on the Moodle platform and on individual audio messages. In this study the efficacy of such online guided pronunciation practice is tested by considering the learners' ability to read with expression, what is sometimes called "prosodic reading" and involves expressive rhythmic and melodic patterns (Dowhower, 1991). Our goal is to determine whether guided online pronunciation practice can lead to improvements in adult EFL learners' ability to read with prosody.

1.2 The training

The pronunciation training within this course targets both segmental and suprasegmental aspects of English pronunciation of Czech speakers. It is grounded in four core assumptions. First, while recognizing the complexity of the relationship between perception and production, we assume that accurate perception is important for developing accurate production (Baese-Berk, 2019; Derwing & Munro, 2015). Consequently, irrespective of the pronunciation feature targeted, the training always includes listening to (multiple and varied) samples of native English speech. The listening tasks guide learners' attention to, and awareness of, specific phonetic features in authentic audio input. Second, it is assumed that building up speech production skills involves proceduralization, i.e. progression from controlled to automatized performance (Gatbonton & Segalowitz, 1988). In each session, initial speaking activities (e.g. imitation, shadowing, or chanting) give the learners opportunities to practise English segmental or prosodic features without overburdening their attentional resources. Subsequently, more demanding tasks (e.g. narrations, rehearsed dramatic dialogues, impromptu role-plays) are introduced. Third, we strive to situate the practice in meaningful contexts to help the learners transfer pronunciation gains to actual language use (Lightbown, 2008). Fourth, we regard feedback as a force that drives learning by helping learners notice the gap between their own pronunciation and that of a native speaker model, thus leading to faster learning gains (Saito & Lyster, 2012).

Although the course had been taught 4 times and received positive evaluations from students, this was the first attempt to empirically test its impact on the learners' output.

The fifth course edition in spring 2020 was organised as follows: the first week's meeting, during which the pre-test data were collected by a research assistant, was followed by 12 training sessions, with post-test data submitted by the learners via email a week after the last session. The training activities addressed multiple aspects of English pronunciation: the perception and production of segmental features typical for the Czech accent in English, as well as phrase- and sentence-level prosody. Learners practiced not only tempo and intonation, the two features of interest for the current paper, but also lexical stress, rhythm, prosodic structuring of discourse, and emotive prosody, amongst other things. Activities targeting tempo and intonation included listening (meaning- and form-focused, of longer and shorter passages, of varied speakers or focused on a single speaker, self- and peer-listening), reading aloud, repetition, imitation, shadowing of varied speech samples, transferring the speech prosody of an example recording to new texts, role-playing based on model recordings, drama rehearsals, rehearsed and impromptu monologue deliveries, recitation of rhythmical verse, and gesture-supported productions. Because of the covid lockdown, the regular 90-minute face-to-face training was replaced by online activities. The course was based firstly on weekly Moodle postings of practice materials, i.e. audio files and handouts, secondly on learners' submissions of self-recordings of all oral tasks, and finally on the instructor's and peers' feedback. Each learner received individualised written and audio feedback on their submissions from the teacher. Selected parts of the learners' recordings were posted on Moodle for peer feedback, which had the form of comprehensibility and accentedness ratings, ratings of the proximity of learners' productions to a native model, or written commentaries on specific pronunciation features. Each week's session was completed by the instructor's feedback on the learner group performance.

1.3 Reading prosody

Good reading prosody reflects readers' consideration of the communicative purpose and it facilitates listeners' comprehension of what is being read. It helps parsing, provides discourse information, directs listeners' attention, adds emphasis, conveys emotions, and offers implicit information. In literature on reading skills of children, reading prosody is sometimes viewed as a component of reading fluency (Kuhn et al., 2010), sometimes fluency and expressivity are separated (Cowie et al. 2002), which is what we also find useful: it is a common experience to encounter an L2 learner who reads aloud fluently but still lacks expressivity in their reading.

Our learners could also read fluently and accurately, as subjectively evaluated by the instructor and one research assistant. They did occasionally produce false starts, self-corrections, hesitations or pauses, as expectable for an adult L2 speaker not trained in reading aloud. They were accurate in the sense of automatically recognizing and instantly producing all the words in the text. On the other hand, it was evident that the learners' ability to read with expressive prosody was rather weak, despite some individual variation. Poor expressive reading is not necessarily only the effect of reading in an L2; adults reading in their L1 also vary in the ability to read with prosody. However, we do think that non-nativeness contributed to reduced expressivity of in most of the learners in this course. Consider the example sentence *"It can't be the milkman because he came this*

morning, and it can't be the boy from the grocer because this isn't the day he comes." Figure 1 shows F0 of a native reading and of a learner pre- and post-training readings. The learner's initial lack of phrasing and pitch dynamics, as compared to native production, is clearly evident.

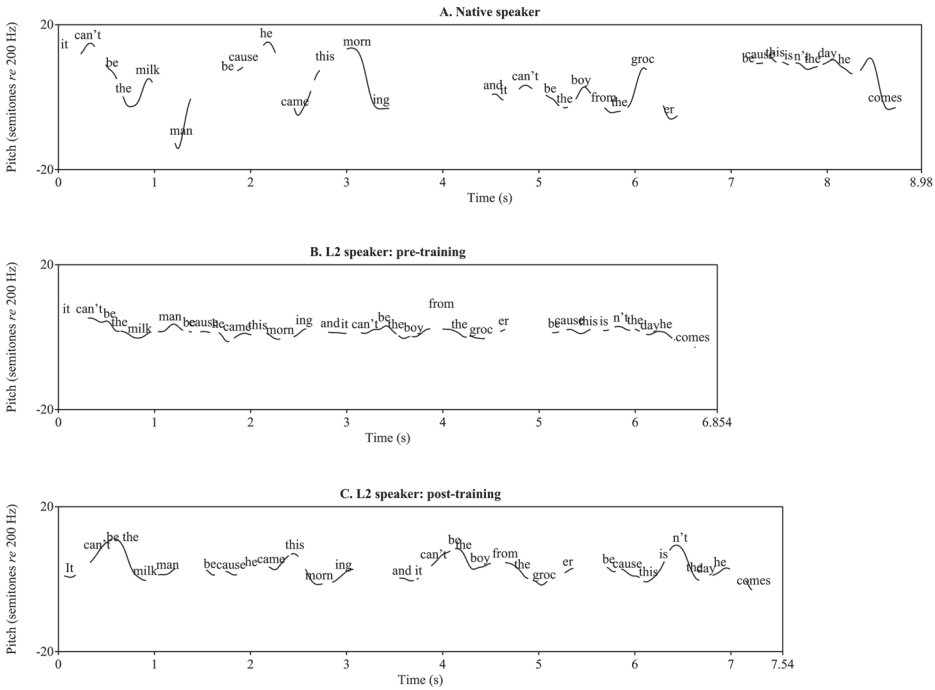

Figure 1. Example of F0 tracks from a native speaker and a learner.

1.4 Correlates of expressivity in this study

We focus on two prosodic correlates of expressivity that have been found relevant in research on the development of reading prosody in children (e.g. Cowie et al., 2002). One is F0 variation measured as the range from the 90th to the 10th percentile. In a subset of data, the magnitude of F0 movements in syllables bearing the nuclear pitch accent was also measured, in other words the difference between the F0 at the beginning of a fall or a rise and the F0 at the end. We expected an increase in expressivity in the post-test data that would be reflected in a wider pitch range.

The other correlate of expressivity considered here is reading tempo. We expected an increase in expressivity to be reflected in *slower* rather than faster reading. While slower speaking rate is often a marker of foreign accentedness (Munro & Derwing, 1998) and even advanced L2 learners are found to speak more slowly than native speakers (e.g. Huang & Gráf, 2020), native listeners may also judge non-native speech as too fast (Munro & Derwing, 2001). What is more, our study is not concerned with the overall speaking tempo but

with a tempo that is appropriate for a specific task, namely for reading with expression. The instruction for the participants was to deliver a story as if for a young audience, in a way that should be engaging. We assume that fast reading tempo in this context actually reduces expressiveness and contributes to monotony. To illustrate the difference between native and non-native reading tempo, in Figure 1 we compare the duration of two renditions of an example target sentence: the native speaker's and an example learner's pre-training rendition. The learner's realisation is approximately 2.3s shorter, which is roughly the time it takes to produce 8-10 syllables. The difference is not only due to the more dramatic pauses evident in the native speaker's rendition. For example, the duration of the clause "*because he came this morning*," pronounced in the example as a fluent intonational phrase both by the learner and the native reader, is 1.43s and 1.74s respectively.

2. Methodology

2.1 Participants

Altogether 16 participants, aged 20 to 27 (mean 22), completed the pronunciation course. The 11 women and 5 men, all native speakers of Czech, were students majoring in English at Palacký University Olomouc. Their general proficiency level was relatively high, between C1 and C2 in CEFR (Council of Europe, 2001): 9 were undergraduate students and had passed the required C1 exam, the remaining 7 were graduate students preparing for the final C2 exam. These participants will be referred to as the Trained Group. Their performance was compared to that of the Comparison group of 14 participants, 9 women and 5 men drawn from the same learner population. They were 12 undergraduate and 2 graduate students, who attended the Foreign Accent seminar, a theoretical course dealing with linguistic, psycholinguistic and sociolinguistic aspects of foreign accents in English.

2.2 Recordings

The test consisted of reading the children's classic story "The tiger who came to tea" (Kerr, & McEwan, 2006). For the Trained Group, the pre-test took place during the first course meeting at Palacký University. The data were collected individually in a sound-proof booth by a student assistant on a Zoom H4n recorder recording the speech at 16-bit and 44.1 kHz. Because of the Covid-19 restrictions, the post-test data were recorded by the learners on their mobile phones (M4A format) in quiet conditions and were generally of a relatively high quality. The Comparison Group data, collected in the spring 2021, were studio-recorded on both tests.

Both on the pre- and the post-test, the participants were instructed to read the story aloud in an engaging way to an imagined audience of pre-schoolers who would view it with pictures on YouTube. The learners were encouraged to rehearse reading the story from the actual picture book (Kerr & McEwan, 2006).

In addition to the Czech learners' data, we analysed the recordings from 7 native English speakers (5 women, 2 men) available on YouTube (see Online data resources). Six of the speakers were judged to be speakers of Southern British English (by a native

speaker of that accent and by a Czech phonetician). One female speaker's pronunciation had features of a northern English accent. The analysis focused on the within-learner pre-vs-post-test improvements. The native speakers' data were used for baseline reference (see Figures 4 and 6).

2.3 Analysis

Altogether 16 intonational-phrase-size units, listed in Table 1, were analysed in terms of reading tempo and F0. Due to an error during the pre-test administration, only the first 12 phrases were available for the analysis in 2 participants' recordings in the Trained group. The complete dataset consisted of 944 phrases; 492 and 448 from the Trained and Comparison groups, respectively.

The phrases were all direct speech statements uttered in the story by 3 different characters – Mommy, Tiger, and Daddy. Utterances were coded as fluent or as involving a disfluency such as a perceptible hesitation or restarting a word. For each fluent utterance, the reading tempo was computed by dividing the number of spoken syllables in a phrase by the total time to read the phrase. The disfluent utterances were excluded from the analysis of tempo. In total 57 utterances were identified as disfluent (pre-test: 24 in the Trained and 13 in the Comparison group; post-test: 5 in the Trained and 15 in the Comparison group). Furthermore, F0 (in semitones *re* 200Hz) was estimated using Praat's autocorrelation algorithm, with all parameters set to default values except the 'Pitch floor' and 'Pitch ceiling', which were set to 60 and 75 Hz, and 500 and 600 Hz for male and female speakers respectively. Since the F0 tracking occasionally results in unreliable outlier values, rather than expressing the F0 range as the measured maximum-minimum, we computed the 80-percentile range (i.e. 90th – 10th percentile). Prior to the analysis, F0 tracks were turned into PraatPitchTier objects, visually inspected and corrected manually for errors such as octave jumps or creaky voice. One utterance out of the 944 was whispered and so it was excluded along with its post-test counterpart.

For the Trained Group, a subset of 7 phrases (italicised in Table 1) was also analysed for the range of the nuclear pitch accent. These phrases were expected to be realised with a falling melody by the learners. Due to an error during the pre-test administration, only 4 phrases were available for the analysis in 2 speakers in the Trained group. A total of 212 phrases were included in this analysis. To analyse the pitch range of the nuclear accent, the most prominent (accented) syllable was identified in each intonational phrase. If the prominent syllable had a clear local F0 peak, the peak was annotated as the F0 maximum. In the following unaccented syllable, the F0 minimum was chosen. If no clear F0 peak could be located and the F0 contour was flat, the stressed syllable in the word most likely to be in focus was marked as bearing the F0 maximum. Figure 2 illustrates the annotation with the phrase "*I think I'd better go now*," realised by the same learner on the pre-test (A) and the post-test (B). In (A), the prominent syllable "*go*" is marked by the local F0 peak while in (B), no clear F0 peak could be identified. The F0 maximum and F0 minimum were sometimes realised within the span of the same syllable when the pitch accent was assigned to the last monosyllabic word (phrases 8, 11, 13, and 14). In 3 recordings the prominent focus was impossible to determine and so they, and their counterparts from the pre- or post-test, were excluded.

Statistical analyses were conducted, and the plots were made in R (version 4.2.0, R Core Team, 2022) using the packages *lme4* (version 1.1.29, Bates et al. 2015), *ggeffects* (version 1.1.2, Lüdecke, 2018), *ggplot2* (version 3.3.6, Wickham, 2016) and *afex* (version 1.1.1, Singmann et al., 2022).

Table 1. The Stimulus Phrases and their Number of Syllables.

Phrase	Sylls.	Phrase	Sylls.
1. I wonder who that can be.	7	9. Excuse me, …	3
2. It can't be the milkman, …	6	10. … *but I am very hungry.*	7
3. … because he came this morning.	7	11. *Thank you for my nice tea.*	6
4. And it can't be the boy from the grocer, …	10	12. *I think I'd better go now.*	7
5. … because this isn't the day he comes.	9	13. *I don't know what to do.*	6
6. And it can't be daddy,	6	14. *The tiger has eaten it all.*	8
7. … because he's got his keys.	6	15. I know what we'll do.	5
8. *We'd better open the door and see.*	9	16. *I've got a very good idea.*	9

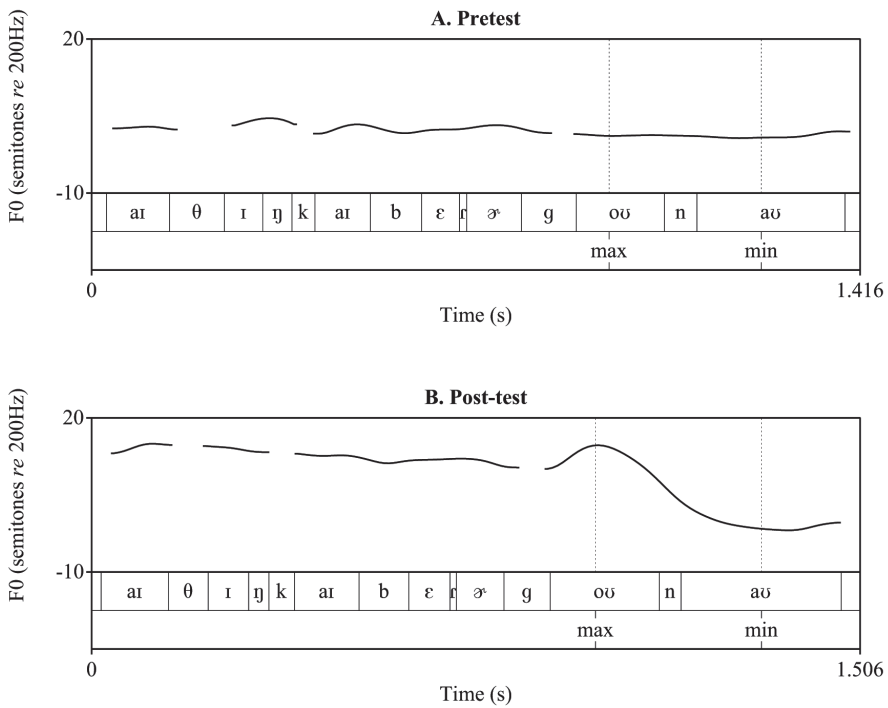

Figure 2. Example Annotation of the Pitch Range (F0 maximum and F0 minimum) of the Nuclear Accent

3. Results

3.1 Reading tempo

Figure 3 shows the distribution of the reading tempo values, expressed as syllables per second, across stimulus phrases and split by participant Group and testing Time. The figure suggests that while the speakers in the Comparison group hardly changed their reading tempo, there seems to have been a slowing down from Time 1 to Time 2 for many speakers in the Trained group.

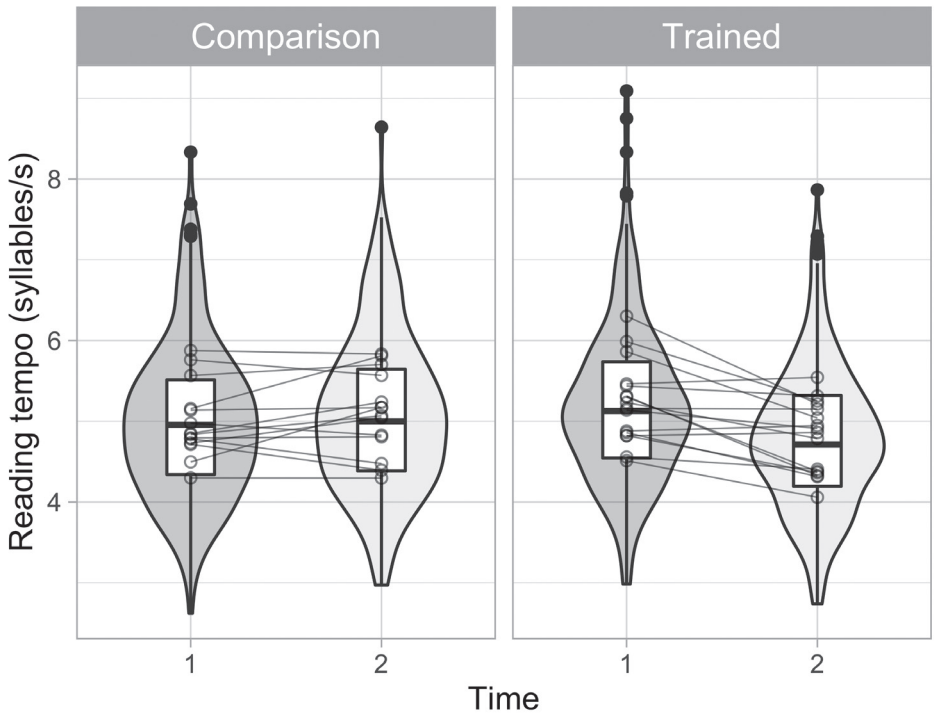

Figure 3. Reading Tempo at Time 1 and Time 2. Violin plots and boxplots show the reading tempo values (in syllables/s) across phrases split by Group and testing Time. Unfilled circles connected by lines show each speaker's means per Time.

To assess whether this difference was reliable, we fitted to the reading tempo data a linear mixed model with testing Time (Time 1 coded as 0, Time 2 as 1), Group (Comparison coded as 0, Trained as 1) and their interaction as fixed effects, and Participant and Phrase as random effects, each with varying intercepts and slopes for Time. Table 2 gives the estimated coefficients and Figure 4 plots the predicted reading tempo values. The model confirmed that whereas the tempo did not change reliably for the Comparison group (Time2 slope +0.09 syllables/s, $p > 0.3$), it did change for the Trained participants,

with a predicted decrease of tempo from Time 1 to Time 2 by 0.09 – 0.49 = -0.4 syllables/s (*SE* = 0.12, *t* = -4.03, *p* < 0.001).

Table 2. Coefficients Estimated by a Linear Mixed Model Fitted to the Reading Tempo Data.

	Estimate	*Std. Error*	*df*	*t value*	*Pr(>|t|)*
(Intercept)	4.9975	0.2156	30.7738	23.1809	**<0.0001**
Time2	0.0899	0.0921	29.8206	0.9766	0.3366
Group2	0.2195	0.1730	29.0724	1.2686	0.2147
Time2:Group2	–0.4915	0.1220	29.2039	–4.0289	**0.0004**

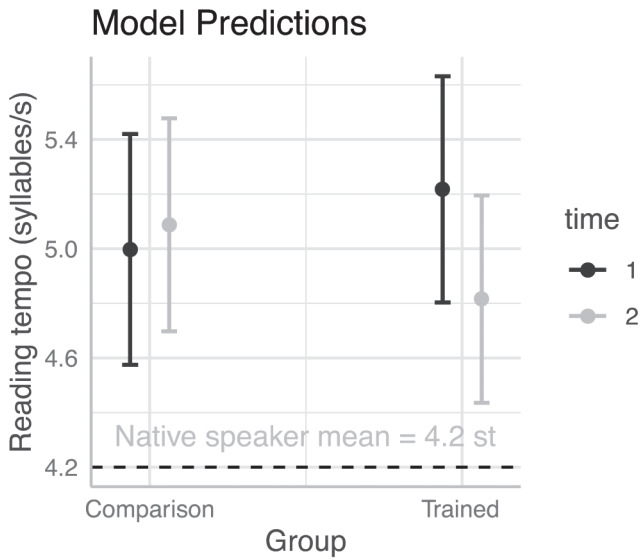

Figure 4. Predicted Reading Tempo Values with a Native-speaker Reference.

3.2 Pitch movement magnitude

As a measure of the magnitude of pitch movement, we used the F0 difference between the 90th and 10th percentile (see 2.3). The values measured, split by Time and Group, are plotted in Figure 5. The plot suggests that whereas there was no change in F0 range in the Comparison group between the testing times, for the Trained group there seems to have been an increase.

To determine whether this is a statistically reliable difference, we fitted to the data another linear mixed model, again with testing Time, Group (both again treatment-coded) and their interaction as fixed effects, and Participant and Phrase as random effects,

each with varying intercepts and slopes for Time. Table 3 gives the estimated coefficients and Figure 6 plots the predicted 80-percentile F0 range values. The model confirmed that while the F0 range did not change significantly for the Comparison group (Time2 slope -0.13 semitones, $p > 0.8$), it did change for the Trained participants, with a predicted increase of F0 range from Time 1 to Time 2 by -0.133 + 2.736 2.6 semitones ($SE = 0.77$, $t = 3.53$, $p = 0.0013$).

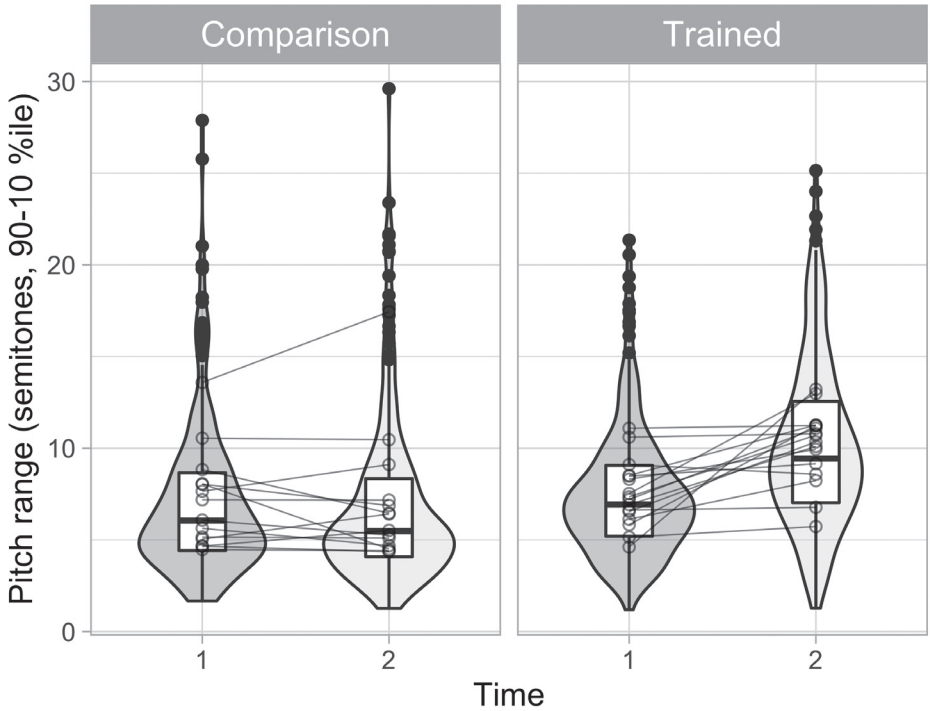

Figure 5. Pitch Range at Time 1 and Time 2. Violin plots and boxplots show the pitch range, measured as the difference of the 90th and 10th percentiles in semitones (re 200Hz), across phrases split by Group and testing Time. Unfilled circles connected by lines show each speaker's means per Time.

Table 3. Coefficients Estimated by a Linear Mixed Model Fitted to the 80-percentile F0 Range Data.

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	7.1205	0.6431	38.4755	11.0715	**<0.0001**
Time2	–0.1332	0.5724	29.5157	–0.2327	0.8176
Group2	0.3881	0.7892	29.9104	0.4918	0.6265
Time2:Group2	2.7361	0.7733	30.0927	3.5381	**0.0013**

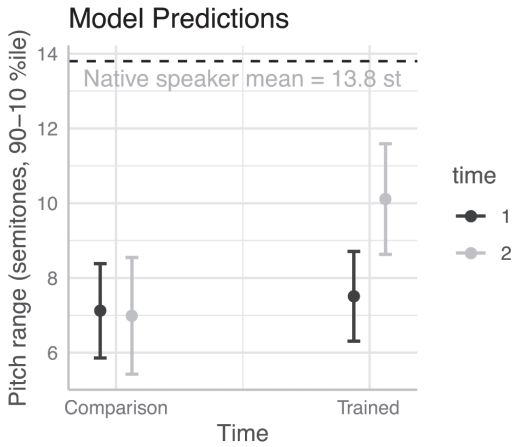

Figure 6. Predicted 80-percentile F0 range values with a Native-speaker Reference.

Next, for the Trained Group we measured the magnitude of F0 movement within nuclear accent contours (see 2.3 above for measurement details). Figure 7 plots the distribution of tonic accent range values for the two testing times, clearly suggesting an increase between times.

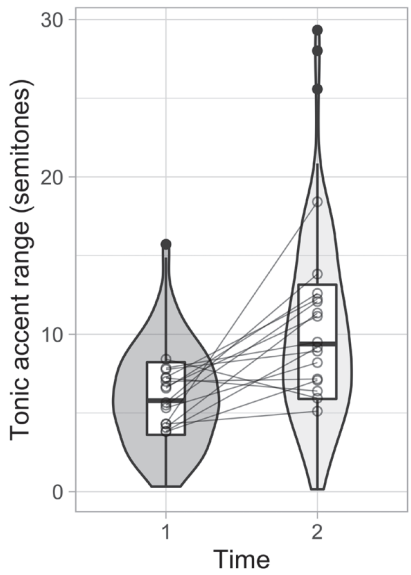

Figure 7. The Nuclear Pitch Accent Range. Violin plots and boxplots show the nuclear pitch accent range, measured as the difference in semitones (re 200Hz) between the F0 maximum in the accented syllable and the F0 minimum in the following unaccented syllable (see more details above in 2.3), across phrases split by testing Time. Unfilled circles connected by lines show each speaker's means per Time.

A linear mixed model fitted to the nuclear accent range data, with testing Time as the fixed effect and Participant and Phrase as random effects, each with varying intercepts and slopes for Time, did not converge. Thus, we refitted this model by restricted maximum likelihood (REML), along with a second model, again by REML, with a reduced random effect structure (see Matuschek et al. 2017 for a criterion for selecting random effects structure supported by the data), namely excluding the correlation of by-Phrase varying intercepts and slopes. A likelihood ratio test (LRT) found no significant difference between the two models ($p > 0.99$), indicating that dropping the correlation parameter did not significantly decrease likelihood. However, when refitting the second model by maximum likelihood, the model again failed to converge. We repeated the procedure, dropping the by-Phrase varying slopes in a third model, resulting in no significant reduction of goodness-of-fit as shown by a LRT ($p > 0.5$), but when the third model was refitted by maximum likelihood, a singularity issue occurred: a correlation between by-Participant varying intercepts and slopes equal to -1 was reported in the output. Thus in a fourth model, the correlation between by-Participant intercepts and slopes was excluded, leading to no decrease in likelihood ($p > 0.99$) but still the output reported a singular fit. A fifth model then removed the by-Participant slopes, this time leading to a significant decrease in likelihood ($p < 0.0001$). Therefore, we refit the fourth model by maximum likelihood, and we report the estimated coefficients in Table 4 (the note recapitulates the final random effect structure). This model on the data from the Trained group predicted the tonic accent range in Time 2 to be about 4 semitones higher than in Time 1 (SE = 0.91, t = 4.30, p = 0.0003).

Table 4. Coefficients Estimated by a Linear Mixed Model Fitted to the Tonic Accent F0 Range Data. Coefficients estimated by a linear mixed model fitted to the nuclear accent F0 range data with Time as the fixed effect and Participant (with varying intercepts and slopes, without slope/intercept correlation) and Phrase (with varying intercepts) as random effects.

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	6.0553	0.4918	14.299	12.313	**<0.0001**
Time2	3.9526	0.9196	23.1834	4.2981	**0.0003**

Since the three models on reading tempo, F0 range, and nuclear pitch accent range, as reported in this section, were fitted to data collected from the same participants, it was necessary to adjust the alpha level from 0.05 to 0.05/3 0.016 (Bonferroni correction). All the p values reported as significant above are lower than that.

4. Conclusion

We examined the effectiveness of a semester-long general pronunciation course in improving prosodic skills of adult EFL learners. The course methodology is based on enhancing both perception and production skills, shifting zoom between phonetic detail

and communicative effectiveness, and engaging learners in working with peer-to-peer as well as instructor-to-learner feedback. The study tested the utility of this methodology in the context of distance learning. The pre-to-post-test comparison focused on the learners' ability to read with prosody, specifically with the appropriate tempo and a wide pitch range. For advanced L2 learners who use English in professional capacities of teachers or interpreters, this is a useful skill.

Regarding the reading tempo, we expected that improvement in expressive reading would be reflected in a slower reading pace. This is what we saw in the post-test, though not uniformly for all trained learners. Closer inspection of the data suggests that duration gains are especially noticeable in the time given to words in focus; a future study should therefore include duration measurement of the pitch movement in the nuclear accented syllable. We also noted the number of disfluent utterances to drop from 24 to 5 in the Trained group, while remaining roughly the same in the Comparison group. It seems that the learners gained fluency as well as expressivity.

The learners also benefited from the training in terms of pitch movement. Clearly, the narrow pitch range regarded as typical for Czech-accented English can be expanded by guided practice even with adult learners who do not have the benefit of authentic input in an English-speaking environment. We measured improvements in the production sentential focus: the nuclear accent had a wider F0 movement in the Trained group's post-test. However, the current study does not consider other aspects of the F0 contour, e.g. pre-nuclear pitch accent, and the alignment of accents and segments.

We can conclude that the course helped the learners improve their L2 prosody in a reading aloud task. Like the example learner in Figure 1, most participants read more slowly and varied their pitch after the training. This was the case despite the training taking place online. While in-class learning has obvious advantages, as learners participate in actual rather than imagined interactions, receiving immediate rather than delayed feedback, the distant training did translate into pronunciation gains. One feature of the online course edition deserves to be explored further: the impact of the learners processing self- and peer-recordings. Also, a future study should compare pronunciation gains from the course taught in the face-to-face vs the distant learning environment to evaluate the importance of genuine oral interaction.

Acknowledgement

This research was supported the Czech Science Foundation grant number 21-09797S.

REFERENCES

Baese-Berk, M. M. (2019). Interactions between speech perception and production during learning of novel phonemic categories. *Attention, Perception, & Psychophysics*, *81*(4), 981–1005.

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1-48. doi:10.18637/jss.v067.i01.

Cowie, R., Douglas-Cowie, E., & Wichmann, A. (2002). Prosodic characteristics of skilled reading: Fluency and expressiveness in 8-10-year-old readers. *Language and Speech*, *45*(1), 47–82.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment assessment*. Cambridge: Cambridge University Press.
De Mareüil, P. B., & Vieru-Dimulescu, B. (2006). The contribution of prosody to the perception of foreign accent. *Phonetica, 63*(4), 247–267.
Derwing, T. M., & Munro, M. J. (2015). *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research*. John Benjamins Publishing Company.
Dowhower, S. L. (1991). Speaking of prosody: Fluency's unattended bedfellow. *Theory into practice, 30*(3), 165–175.
Gatbonton, E., & Segalowitz, N. (1988). Creative automatization: Principles for promoting fluency within a communicative framework. *TESOL quarterly*, *22*(3), 473–492.
Hirschberg, J. (2002). Communication and prosody: Functional aspects of prosody. *Speech Communication*, *36*(1-2), 31–43.
Huang, L. F., & Gráf, T. (2020). Speech Rate and Pausing in English: Comparing Learners at Different Levels of Proficiency with Native Speakers. *Taiwan Journal of TESOL*, *17*(1), 57–86.
Kang, O. (2008). *Ratings of L2 oral performance in English: Relative impact of rater characteristics and acoustic measures of accentedness* (Doctoral dissertation, University of Georgia).
Kang, O., Rubin, D. O. N., & Pickering, L. (2010). Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English. *The Modern Language Journal*, *94*(4), 554–566.
Kerr, J., & McEwan, G. (2006). *The tiger who came to tea*. HarperCollins Childrens' Books.
Kuhn, M. R., Schwanenflugel, P. J., Elizabeth B. Meisinger, E. B. 2010. Aligning Theory and Assessment of Reading Fluency: Automaticity, Prosody, and Definitions of Fluency. *Reading Research Quarterly 45*(2): 232–253.
Lengeris, A. (2012). Prosody and second language teaching: Lessons from L2 speech perception and production research. In J. Romero-Trillo (Ed.) *Pragmatics and prosody in English language teaching*, Springer. 25–40.
Lightbown, P. M. (2008). Transfer appropriate processing as a model for classroom second language acquisition. In Z. Han (Ed.), *Understanding second language process* (27–44). Multilingual Matters.
Lüdecke, D. (2018). ggeffects: Tidy Data Frames of Marginal Effects from Regression Models. *Journal of Open Source Software*, 3(26), 772. doi:10.21105/joss.00772.
Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language, 94,* 305–315.
Mennen, I., & de Leeuw, E. (2014). Beyond segments: Prosody in SLA. *Studies in Second Language Acquisition*, *36*(2), 183–194.
Munro, M. J., & Derwing, T. M. (1998). The effects of speaking rate on listener evaluations of native and foreign-accented speech. *Language Learning*, *48*(2), 159–182.
Munro, M. J., & Derwing, T. M. (2001). Modelling perceptions of the accentedness and comprehensibility of L2 speech the role of speaking rate. *Studies in second language acquisition*, *23*(4), 451–468.
R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
Saito, K., & Lyster, R. (2012). Effects of form-focused instruction and corrective feedback on L2 pronunciation development of/ɹ/by Japanese learners of English. *Language learning*, *62*(2), 595–633.
Singmann H., Bolker B., Westfall J., Aust F., & Ben-Shachar M. (2022). *afex: Analysis of Factorial Experiments*. R package version 1.1.1, https://CRAN.R-project.org/package=afex.
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.
Volín, J., & Skarnitzl, R. (2010). Suprasegmental Acoustic Cues of Foreignness in Czech English. *Proceedings of Odyssey: The Speaker and Language Recognition Workshop Odyssey 2010*. Brno: VUT, 271–278.
Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag.

Online data resources

YouTube. (2020, May). https://www.youtube.com/watch?v=UGqFxN0V-CA
YouTube. (2020, May). https://www.youtube.com/watch?v=6Cba9CUQUD0
YouTube. (2020, May). https://www.youtube.com/watch?v=IUAr_bnXtKc (no longer accessible)
YouTube. (2020, May). https://www.youtube.com/watch?v=Lok2QdctuCg (no longer accessible)

YouTube. (2021, June). https://www.youtube.com/watch?v=4k87lP1GNrE
YouTube. (2021, June). https://www.youtube.com/watch?v=ggjemuhlmuM
YouTube. (2021, June). https://www.youtube.com/watch?v=RjcxZemAxNI

Šárka Šimáčková
Dept. of English & American Studies
Palacký University Olomouc
Olomouc, Czech Republic
E-mail: sarka.simackova@upol.cz

Václav Jonáš Podlipský
Dept. of English & American Studies
Palacký University Olomouc
Olomouc, Czech Republic
E-mail: vaclav.j.podlipsky@upol.cz

2022 ACTA UNIVERSITATIS CAROLINAE PAG. 131–149
PHILOLOGICA 1 / PHONETICA PRAGENSIA

LONGITUDINAL STUDY OF PHONETIC DRIFT IN L1 SPEECH OF LATE CZECH-FRENCH BILINGUALS

MARIE HÉVROVÁ, TOMÁŠ BOŘIL

ABSTRACT

This study investigates temporal development of phonetic drift (i.e., when L1 pronunciation is affected by acquiring an L2 language) in the L1 speech of four Czech university students (two female and two male) who went to study in Toulouse as part of the Erasmus programme. Having started studying L2 French at the age of twelve to sixteen, they are considered the so-called Czech-French late bilinguals. The subjects were recorded reading out a Czech text and producing semi-spontaneous speech in three sessions – immediately after their arrival, and then at the end of the first and the third month of their stay in France. Based on acoustic analyses, we statistically evaluated the formant frequencies of vowels, the spectral moments of the fricatives /ɦ/ and /x/, and the production frequency of schwa in the word-final position, which is a distinctive pronunciation feature for Toulouse French. Even though speech and its development are highly individual, we were able to witness certain pronunciation shifts regarding all the examined phones. However, the majority of statistically significant shifts were linked to the formant values of vowels.

Keywords: phonetic drift, late Czech-French bilinguals, vowel quality, spectral moments, word-final schwa

1. Introduction

The influence of the first language (L1) of an adult speaker on the acquisition of the second language (L2) has been studied extensively at the phonetic level (see Aoyama & Guion, 2007; Colantoni & Steele, 2007; Curtin, Goad & Pater, 1998; Holliday, 2015; Kijak, 2009; Major, 1986, among many others). However, the influence of L2 on L1 of an adult speaker, who started to learn the L2 after the age of six and thus is considered a late-bilingual speaker, is a topic explored by fewer recent studies which typically deal only with partial issues. The majority of them compare the L1 of monolinguals with the L1 of late-bilinguals living in an L2 country for a couple of years, both recorded once at a specific time (see, e.g., Bergmann, Nota, Sprenger & Schmid, 2016; De Leeuw, 2008; Kupske & Alves, 2016; Lang & Davidson, 2019; Major, 1992; Mayr, Price & Mennen, 2012; Stoehr, Benders, van Hell & Fikkert, 2017; Sůčková, 2020; Ulbrich & Ordin, 2014).

https://doi.org/10.14712/24646830.2022.33

However, longitudinal studies investigating the evolution of L1 of the late bilinguals are rare (see section 1.1 below).

The existence of several differences between the Czech and French languages, both at segmental and suprasegmental levels (Hévrová, Bořil & Köpke, 2020; Hévrová, 2021; Paillereau, 2015; Skarnitzl, Šturm & Volín, 2016), encouraged Hévrová (2021) to suppose that L1 of Czechs living in Southern-French Toulouse and its surroundings will be influenced by their everyday use of French. The comparison of their L1 with the L1 of Czech monolinguals supported the hypothesis. However, in studies based on the comparison of a group of L1 late-bilinguals with another group of monolinguals, it is complicated to distinguish whether the differences between their L1s exist only due to the effect of moving to the L2 country or whether it was already present before (Hévrová, 2021). To deal with such issues, this paper features a longitudinal study capturing a gradual evolution of L1 of a speaker moving to an L2 country.

1.1 Longitudinal studies of L2 influence on L1

The effect of an L2 influence on the L1 is often referred to by a wide range of terms where three of them are the most common (Gallo et al., 2021; Köpke, 2004): a first language attrition, a cross-linguistic influence (CLI) and a phonetic drift. The first language attrition is commonly associated with a non-pathological and non-ageing effect of changes in L1 of a late bilingual resulting from a long-term immersion into an L2 environment (Köpke & Schmid, 2004; Kornder & Mennen, 2021). These changes are linked to a decreased L1 use and input (cf. De Leeuw, 2019) and are considered to be "long-term L1 changes", according to Chang (2019, p. 192). Contrarily, the phonetic drift refers to "ostensibly short-term changes" in bilinguals' L1 speech resulting from "recent L2 experience" (Chang, 2019, p. 192) and L2 "exposure" (Tobin, Nam & Fowler, 2017, p. 46). A phonetic drift is linked with "cases of a subtle phonological restructuring in the L1" (Chang, 2012, p. 249). Finally, the term CLI introduced by Sharwood Smith (1983) means any influence of one of a speaker's languages on another (cf. Jarvis & Pavlenko, 2008; Pavlenko, 2000). In the present study, we will be examining phonetic drift as this term best captures the nature of L2 influence on L1 of our bilingual respondent similarly to other studies (e.g., Chang, 2012; Tobin et al., 2017).

Speech Learning Model (SLM) (Flege, 1995) and its revised version SLM-r (Flege & Bohn, 2021) are widely employed in studies of phonetic L2 influence on L1 (see, e.g., Bergmann et al., 2016; Chang, 2010; De Leeuw, 2008; Hévrová, 2021; Kornder & Mennen, 2021; Lang & Davidson, 2019; Mayr et al., 2012; Sůčková, 2020). The key suppositions of SLM is that L1 and L2 exist in a common phonetic space and interact with each other (Flege, 1995; Flege & Bohn, 2021), which may lead not only to a non-native L2 speech production but also to a less native-like L1 production (see, e.g., Sancier & Fowler, 1997; De Leeuw, Schmid & Mennen, 2010; Bergmann et al., 2016; Mayr, Sánchez & Mennen, 2020; Hévrová, 2021). Results of acoustic analyses of the less native-like L1 production may be interpreted as the assimilation or dissimilation effect (see De Leeuw, 2019) according to the type of changes occurring in L1 phones. Assimilation is a shift of an L1 sound towards an L2 category, while dissimilation refers to the speaker's effort to maintain a difference between L1 and L2 sound, leading

to greater acoustic distance between these two sounds (De Leeuw, 2019). Both assimilation and dissimilation effects are commonly linked with a phonetic drift (Chang, 2012; Kartushina, Hervais-Adelman, Frauenfelder & Golestani, 2016), and the type and extent of a change vary widely with individual bilinguals (Bergmann et al., 2016; De Leeuw, Tusha & Schmid, 2018; Major, 1992; Mayr et al., 2012). For instance, in De Leeuw's (2008) analysis of the tonal alignment of prenuclear rise, two of ten late German-English bilinguals exceeded the monolingual German norm at the end of the rise in their L1, thus evincing a dissimilation instead of expected assimilation; in the remaining eight bilinguals assimilation was confirmed. These results follow the SLM-r supposition that L2 sound production and perception do not solely depend on the phonetic systems of L1 and L2 of the bilinguals but also on many endogenous factors which may vary within an individual bilingual and thus cause the differences in organisation and interaction between L1 and L2 phonetic categories in the bilingual's common phonetic space.

According to the Second Language Linguistic Perception model (L2LP) (see Van Leussen and Escudero (2015) for the revised version), in the final state of L2 learning, L2 learners separate L1 and L2 grammars and language activation modes that allow them to attain optimal perception of L2 and preserve the one of L1 (Escudero, 2005). To maintain the optimal L1 and L2 perception and production, learners must be exposed to rich L1 and L2 input, otherwise L2 will affect L1 (Elvin and Escudero, 2019).

Longitudinal studies examining phonetic drift in L1 of late bilinguals during a short stay in an L2 country are rare; the majority of studies focus on L2 English, in which participants are often considered as early bilinguals due to their age of L2 acquisition. Chang (2012) focused on L1 of 36 American English learners of Korean. From this sample, a group of 19 "functionally monolingual" learners were selected (3 males and 16 females) and enrolled in a 6-week Korean language course of 4 hours per weekday at the South Korean university. The participants reading a word list in L1 were recorded in five instances after each week of the course. Nine native Korean monolingual speakers represented a control group with the same reading task. Significant changes in the first formant (F1) values of L1 vowels produced by female learners after five weeks of the course were found; the size of the male group was insufficient for statistical significance. The drift was consistently unidirectional for all vowels in general in accordance with the mutual position of Korean and English vocalic systems instead of assimilation of individual's L1 vowels. Mayr et al. (2012) also found assimilation in F1 of the whole L1 vocalic system in the study of phonetic attrition, while this was not observed in Hévrová (2021).

Kartushina et al. (2016) showed that phonetic drift could appear very quickly, i.e., after one hour of intensive training of target foreign vowels. Interestingly, five weeks of intensive L2 courses and staying in an L2 environment sufficed for phonetic drift appearance in the work of Chang (2010, 2012) but not in the study of Lang and Davidson (2019). The discrepancy between the two studies might be related to external factors such as the number of hours of L2 learning classes or characteristics of the vowel system of each language. Nevertheless, because of these differences and variability of individual speakers as described by SLM-r, the time duration of contact with the L2 to cause the phonetic drift cannot be precisely determined.

In Chang (2013), the phonetic drift was stronger in novice learners (learners with no prior knowledge of the L2) rather than in experienced learners enrolled in the same language program. In contrast, it occurred more in advanced learners than in beginners (Herd, Walden, Knight & Alexander, 2015) and in a long-term L2 country stay than in a short-term stay (Lang & Davidson, 2019). According to studies by Flege (1987) and Major (1992), the more speakers are proficient in L2, the more drift occurs. Consequently, the result of this study seems to cast doubt on Chang's hypothesis (2013) that the drift gets greater with less experienced speakers.

Looking at the respondents in studies on phonetic drift, most studies consider a few speakers as a group (see, e.g., Chang, 2012) or are focused on a single speaker (see, e.g., Sancier & Fowler, 1997). Moreover, longitudinal studies focusing on more than one speaker analysing a phonetic drift of each speaker individually are rare. For this reason, we have decided to analyse the speech of 4 speakers separately.

1.2 L2 influence on L1 of late Czech-French bilinguals

L2 influence on L1 of late Czech-French bilinguals was examined by Hévrová (2021) both at a segmental and suprasegmental level. Two experiments were conducted in which L1 speech production in a reading aloud task and semi-spontaneous speech of late Czech-French bilinguals, mainly living in Toulouse geographical area, was compared with that of Czech monolinguals. In the first experiment, a perception test revealed that Czech monolingual listeners perceived the bilinguals' semi-spontaneous L1 speech as significantly less typically Czech sounding than Czech monolingual speakers, but this was not the case for the reading task. In the second experiment, the speech of 17 bilinguals and 17 monolinguals was analysed acoustically, and the phonetic cross-linguistic influence (CLI) was mainly found in spectral characteristics of several of the bilinguals' vowels, /ɦ/ and /x/, in the non-conclusive intonation patterns as well as in the frequency of use of schwa in the word-final position. A correlation analysis was performed between phonetic L2 influence on L1 and several extralinguistic factors, such as the frequency of use of L1 by the bilinguals, their length of residence in France, proficiency in L2 and preferences for either L1 or L2 country, culture and language. A significant link of phonetic L2 influence on L1 in /ɦ/ in semi-spontaneous speech and in /x/ in the reading task with the proficiency in French of the late bilinguals was found. Bilingual with a higher proficiency in French showed less phonetic attrition than those with a lower proficiency.

1.3 Acoustic properties of selected elements of Czech and French phonetic system

Standard Czech and Standard French[1] differ in the number of vowels (see Table 1) and their articulatory properties (see Table 2 for the relation of formant values of monophthongs). In Toulouse French (spoken mainly by people born and living in the geograph-

[1] Language varieties preferred in television or radio broadcast and education; geographically typical for the Bohemian and Paris regions respectively.

ical area of Toulouse), some speakers do not distinguish between /e/ and /ɛ/, /a/ and /ɑ/, /o/ and /ɔ/ and /œ/ and /ø/ in their speech production, while others make this distinction or use /e/ and /ɛ/, /a/ and /ɑ/, /o/ and /ɔ/ and /œ/ and /ø/ according to rules different from Standard French (Courdès-Murphy, 2018; Durand, 2009). Most often, nasal vowels in Toulouse French are pronounced as an oral vowel followed by a very short nasalised vowel and a long nasal consonant (Durand 1988; Delvaux, Kathy, Piccaluga & Harmegnies, 2012).

Czech /ɦ/ and /x/ are two fricatives that do not exist in Standard French or Toulouse French. Hévrová (2021) calculated four mean spectral moments (centre of gravity (*COG*), standard deviation, skewness and kurtosis) of /ɦ/ and /x/ from the L1 semi-spontaneous speech and /x/ in a reading aloud task of 17 Czech female monolinguals, data collected by Tykalová et al. (2021) (see Table 3). The spectral moments of Czech /x/ were also measured by Sedláčková (2010) on recordings of read news in Czech Radio 1 – 'Radiožurnál' by 21 Czech moderators (also see Table 3).

Some French speakers may pronounce a schwa at the end of specific words (so-called 'e-muet' – the dumb 'e') (cf. Brun, 2000). At the geographical level, this schwa (hereinafter referred to as final schwa) is rarely pronounced by speakers from Northern France, while it is often pronounced by southern French speakers. The pronunciation of the final schwa is practically systematic in Toulouse French, and its duration is usually longer in the production of Toulouse French speakers than the French from Marseilles (cf. Coquillon, 2005). From the phonetic point of view, the final schwa corresponds to the sound [ə] stuck to the last pronounced consonant of the word or, in very few attested cases, to the last pronounced vowel of the word with a consequence of creating a new syllable (Carton, Rossi, Autesserre & Léon, 1983; Coquillon, 2005; Durand, Slater & Wise, 1987). At the orthographical level, the final schwa may match up with the letter 'e' at the end of the word, but it may also be pronounced even if there is no such letter (cf. Coquillon, 2005). For example, 'mère' may be pronounced as [mɛRə] (see Pustka, 2011), 'alors' as [alɔRə], and 'avec' as [avɛkə] (see Carton et al., 1983).

French speakers may express a hesitation by employing the final schwa, and Candea (2000) proposed to use duration as a parameter for distinguishing a final schwa as a simple indication of the geographical origin of the speaker from a final schwa as an expression of hesitation. In the corpus of production of Standard French speakers, the final schwa labelled as the expression of hesitation lasted from approximately 150 to 500 ms. The final schwa was rarely found in Standard Czech (Průchová, 2016), while a schwa separated from the words by silences is commonly employed for an expression of a hesitation (Šulecová, 2015; Volín, 2010).

Table 1: Vowels of Standard Czech and Standard French. Source: Léon (1997); Volín (2010); Skarnitzl et al. (2016).

		Standard Czech	**Standard French**
Monophthongs	Oral	ɪ, iː, ɛ, ɛː, a, aː, o, oː, u, uː	i, e, ɛ, a, ɑ, u, o, ɔ, y, ø, œ
	Nasal	-	ã, ɛ̃, œ̃, ɔ̃
Diphthongs	Oral	o͡u̯, a͡u̯, ɛ͡u	-

Table 2: Differences among formant values of Standard Czech, Standard French and Southern French vowels. Note: StCZ = Standard Czech, StFR = Standard French, SFR = Southern French. Dark grey colour means the most important difference, grey indicates important difference, light grey means less important difference. According to: Skarnitzl and Volín (2012); Tubach (1989); Paillereau and Chládková (2019); Gendrot and Adda-Decker (2005); Woehrling (2009).

Vowel	F1	F2
/ɪ/ or /i/	StCZ > StFR and SFR	StCZ < StFR and SFR
/ɛ/	StCZ > StFR and SFR	StCZ < StFR and SFR
/a/	StCZ > SF > StFR	StCZ < SF < StFR
/u/	Some differences but not possible to determine precisely	Some differences but not possible to determine precisely
/o/	StCZ > StFR and SFR	Some differences but not possible to determine precisely

Table 3: Four mean spectral moments of Czech fricatives /ɦ/ and /x/. Centre of gravity (*COG*), standard deviation, skewness and kurtosis of /ɦ/ in semi-spontaneous speech and of /x/ in read news, reading task and semi-spontaneous speech.

	***COG* (Hz)**	**St_dev (Hz)**	**Skewness**	**Kurtosis**
/ɦ/ semi-spontaneous (Hévrová, 2021)	337	580	23	876
/x/ read news (Sedláčková, 2010)	1191	1373	3	18.1
/x/ reading task (Hévrová, 2021)	1127	1622	7	86
/x/ semi-spontaneous (Hévrová, 2021)	1199	1654	6	83

1.4 Hypotheses

Concerning the SLM, SLM-r, L2LP, the results of studies on phonetic drift, the CLI found in the L1 speech of late Czech-French bilinguals (Hévrová, 2021) and the phonetic differences between Czech and French, (i) we predict phonetic drift may appear in L1 speech of Czech Erasmus students in Toulouse, more particularly in spectral moments of their /ɦ/ and /x/ (due to the lack of their L1 input in accordance with the L2LP), use of the final schwa in the semi-spontaneous speech and some formants of several vowels, mainly: F1 of /aː/, F1 and F2 of /ɛ/, F1 and F2 of /ɛː/, F1 and F2 of /ɪ/, F1 and F2 of /iː/. With respect to the consideration of individual differences corresponding to SLMr, (ii) we predict a large variability in the type and amount of the drift in the L1 speech of Czech Erasmus students in Toulouse.

2. Method

2.1 Respondents

For the present study, we recorded the L1 speech produced by 5 native Czech students coming from Bohemia of the Czech Republic to Toulouse for Erasmus, living in the Toulouse area during their stay. All respondents claimed that they had not lived in any region with a strong variety of Czech, nor had they spoken with a Moravian accent. They all self-reported having a good English proficiency level (B2 or C1). They all started to learn French during their grammar school (being approximately from 12 to 16 years old) except for one speaker who only attended 4 hours of online French learning one month before coming to Toulouse. This particular respondent scarcely used French at the university in Toulouse since most of their classes were in English; because of these facts, we decided to exclude this speaker from the studies. For the remaining four speakers, see Table 4.

Table 4: Personal data and language background of speakers.

Speaker	Sex	Age	Foreign countries where they lived for more than 6 months	Czech region they lived the longest
LS1	M	33	Poland (13 months)	North Bohemia
LS3	F	21	none	Central Bohemia
LS4	M	27	France – Angers (6 months)	Central Bohemia
LS5	F	20	none	South Bohemia

2.2 Procedure

The L1 speech of each student was recorded at three distinct times: first, when the student arrived in Toulouse, second, after about five weeks of the student's stay in Toulouse and third, after about three months of the student's stay there. Table 5 gives the precise number of days after arrival when the recordings were made. The first author of the article was quickly notified about the exact day of students' arrival to Toulouse. Nevertheless, this was not always possible, and for two students, the initial recording was delayed. The first author of the present article recorded the students' speech production in a quiet recording studio (PETRA) at the University of Toulouse Jean-Jeaurès using a sound card MOTU UltraLite-mk3 and Neumann TLM 49 microphone located around 20 cm from the speakers' mouth.

Table 5: Days after arrival (A) when the recording was made.

Speaker	1st recording	2nd recording	3rd recording
LS1	A+2	A+40	A+90
LS3	A+1	A+35	A+92
LS4	A+15	A+35	A+93
LS5	A+21	A+36	A+87

During each recording session, at first, the participant had to produce one minute and a half of semi-spontaneous speech in French (speaking about one or more proposed or free topics) before starting to accomplish the speech production tasks in Czech. This was performed in order to ensure the authentic environment in which our students lived: speaking French during the day and switching to Czech occasionally. The proposed topics were: plans for holidays or the weekend, typical day, studies, family, hobbies, and others. The first speech production task in Czech consisted of one minute and a half of semi-spontaneous speech in Czech (hereafter SS) on one or more proposed topics similar or identical to those proposed for the production of semi-spontaneous speech in French. The second speech production task consisted of reading aloud a short Czech text (i.e., reading task, henceforth RT). For the first recording session, a short text extracted from Čapek (1960) was used; for the second session, we used a short text from Čapek (1939) and for the third session, a short text called 'Milánek' was employed which is a part of standardised protocol for language and acoustic assessment and analysis. All the texts are frequently used for the research purposes at the Institute of Phonetics of Charles University. The texts were similarly long and easy to read. Other production tasks were also recorded but not used for the present study.

2.3 Acoustic analysis

All recordings were orthographically transcribed into phrase tiers, semi-automatically segmented and labelled into word and phone tiers in Praat (Boersma & Weenink, 2019). The segmentation and labelling were manually corrected following the rules of segmentation (Machač & Skarnitzl, 2009); for instance, the vowels' boundaries were placed according to the presence of full formant structure, and initial glottal stops and final voice decay time were not considered as a part of the vowel. The annotation of the final schwas [əː] was guided by their duration, sticking to the end of the word and a perceptual creation of a new syllable. The schwa separated from the end of the word by a glottal stop [ʔ] was not considered a final schwa.

Four spectral moments of /ɦ/ and /x/ were measured automatically using a Praat script, computing the mean of the given spectral moment from the second third of the vowel duration to minimise the coarticulation's effect on the formant value. The F1 and F2 values of vowels were measured in the middle third of their duration using

the Burg method in Praat (Boersma & Weenink, 2019). Three different settings of the method were used with window size of 25 ms and 50 Hz pre-emphasis in all cases: (i) the maximum number of formants: 5, formant ceiling of 5500 Hz, (ii) max. 5 formants and 3000 Hz ceiling, (iii) max. 10 formants and 3000 Hz ceiling. For each setting and formant, a mean value of estimates was obtained. Then, based on a visual inspection of spectra and an auditory perception, we manually chose the most appropriate values (from the estimates proposed by the three settings) not containing nasal formants, F1 with f_0 merging and other typical estimation errors. In most scenarios, the first two values of the method (i) agreed best with our manual inspection of /ɪ, ɛ, a/ vowels and the first and the third value of the method (ii) performed well with /o, u/ vowels. However, in many situations, this was not a rule. Nasal context, creeky-voice, different f_0 and spectral composition of the voice played a significant role. Hence, the manual evaluation was necessary.

2.4 Statistical analysis

For the statistical analysis, we excluded the phonemes in foreign words such as the names of French or foreign cities, unpronounced and semi-pronounced phones (annotated manually in brackets, e.g., typically the vowel /o/ in the Czech word 'protože'), the Czech conjunction /a/ (meaning "and" in English) longer than 150 ms being considered as a hesitation (cf. Rubovičová, 2014).

The data were analysed in R (R Core Team, 2019) using the packages *lme4* (Bates, Mächler, Bolker & Walker, 2015), *dplyr* (Wickham, François, Henry & Müller, 2020), *rPraat* (Bořil & Skarnitzl, 2016), *ggplot2* (Wickham, 2016), and *emmeans* (Lenth, 2021). For the study of vowels and /ɦ/ and /x/, we first counted the number of their occurrences by speaker, recording session and task, see Table 6. Groups with less than 4 occurrences were excluded from the study (/ɦ/ in RT of all speakers, /ɦ/ in SS of the speaker LS3, and /ɛː/, /oː/, /uː/ in both tasks).

In order to examine the differences in the acoustic properties of vowels, /ɦ/ and /x/ across the three recording sessions for each speaker separately, we performed a set of linear regression models with the interaction between two fixed effects for each studied acoustic property and phoneme. The fixed effects were the recording session (hereafter time) with three levels (1st recording = A0, 2nd recording = A1, 3rd recording = A2) and task (two levels: RT and SS). We analysed the relationship between the recording session and the given acoustic property: *lm*(*value ~ time∗task, data*). Visual inspection of residual plots did not reveal any obvious deviations from homoscedasticity or normality. The comparison of estimated means across levels of the effects was carried out with the package *emmeans* on the full model with the interaction. The significance level of 0.05 with Bonferroni correction for 4 speakers was $\alpha = 0.0125$.

For the study of the final schwa, we counted the number of occurrences by speaker and recording session in SS. During the manual annotation, we did not observe any occurrence in RT, which corresponds to the results of Hévrová (2021). Thus, the final schwa was not analysed in RT.

Table 6: Number of occurrences of analysed phonemes by speaker and task.

	Reading task (RT)				Semi-spontaneous speech (SS)			
	LS1	LS3	LS4	LS5	LS1	LS3	LS4	LS5
/ɦ/	–	–	–	–	21	–	17	26
/x/	22	19	19	17	22	29	18	25
/a/	113	113	106	110	205	272	218	189
/aː/	30	28	29	29	68	70	78	54
/ɛ/	152	154	156	155	327	347	384	301
/ɪ/	79	80	82	83	142	225	173	177
/iː/	64	58	62	62	99	99	99	61
/o/	108	106	107	106	216	284	234	218
/u/	42	43	44	44	47	89	71	67

3. Results

3.1 Spectral moments in /ɦ/ and /x/

The analysis of the four spectral moments (*COG*, standard deviation, skewness and kurtosis) of /ɦ/ in SS for the three speakers (see Figure 1) showed a significant difference

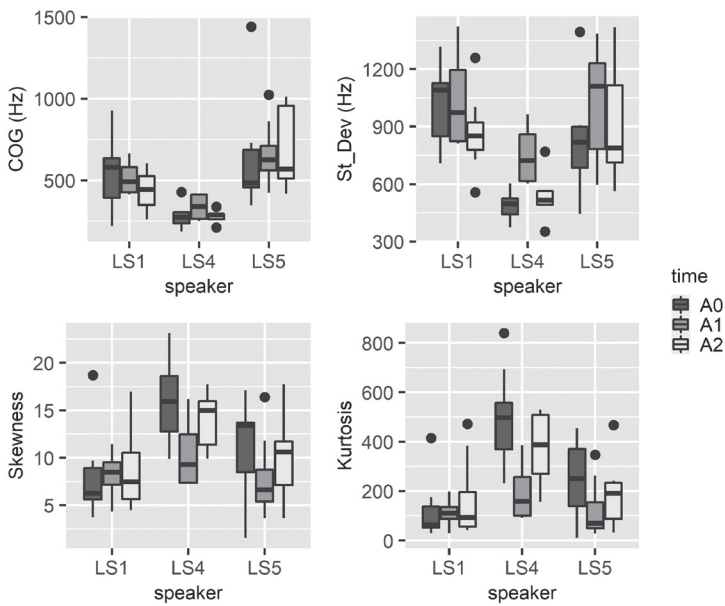

Figure 1: *COG*, standard deviation, skewness and kurtosis of /ɦ/ in semi-spontaneous speech (SS).

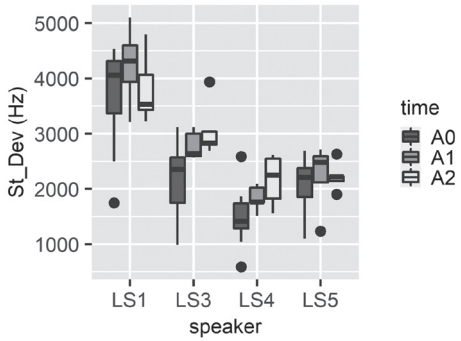

Figure 2: Standard deviation of /x/ in reading task (RT).

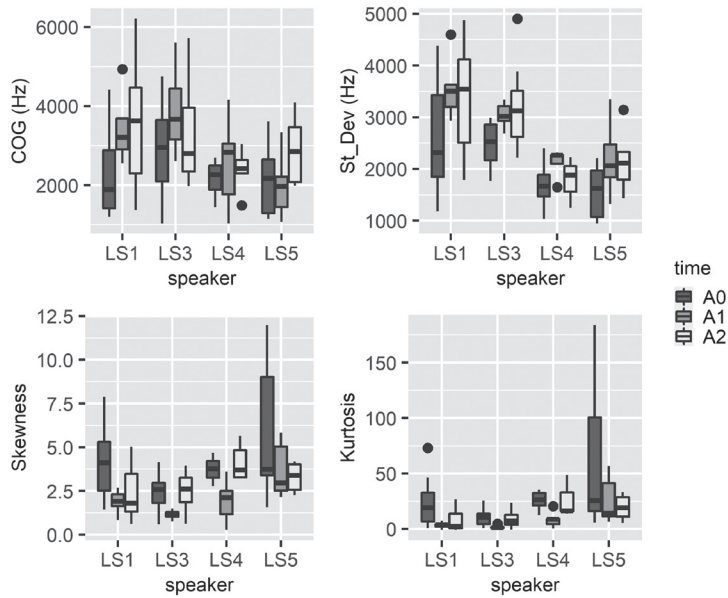

Figure 3: *COG*, standard deviation, skewness and kurtosis of /x/ in semi-spontaneous speech (SS).

in the standard deviation between the first and the second recording session (A1 – A0) of LS4 speaker, estimate: 264.3 Hz (SE = 76.8 Hz, DF = 14, $p = 0.0104$), where SE is the standard error, and DF denotes degrees of freedom.

The analysis of /x/ in RT did not reveal any significant difference in the four spectral moments due to the Bonferroni correction, although there is an indicated upward trend over time in the standard deviation (depicted in the Figure 2).

The analysis of /x/ in SS of all speakers (see Figure 3) lead to a significant difference in skewness between the second and third recording session (A2 – A1) of LS4, estimate: 2.20 (SE = 0.614, DF = 15, $p = 0.0071$).

3.2 F1 and F2 in vowels

Distributions of F1 and F2 formant values during three recording sessions are depicted in Figure 4. All significant shifts between two time moments are summarised in Table 7. In RT, the most frequent are shifts in /a, aː/ where a significant shift in one of the formants is present in all speakers. In contrast, shifts are rare in the remaining vowels in RT. In SS, significant shifts across all analysed vocals /a, aː, ɛ, ɪ, o, u/ are more frequent. In relation to Figure 4, the individuality of each speakers is notable. Please also see Table 9 in the Discussion and conclusions section for a different way of illustrating the significant shifts found.

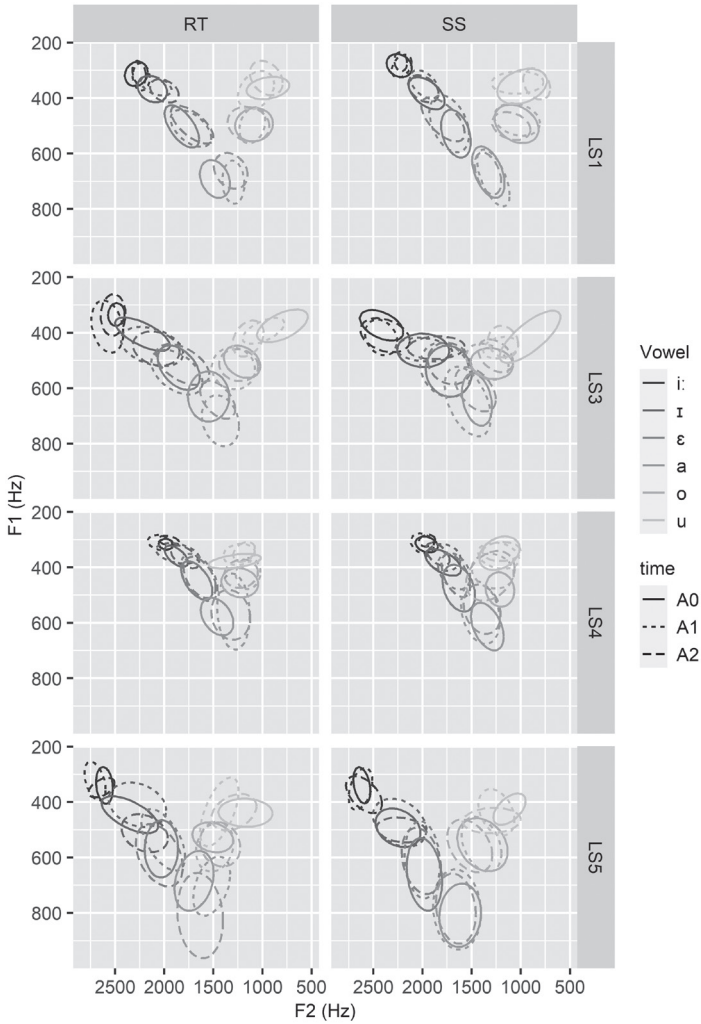

Figure 4: Formant values (50%) of vowels in reading task (RT) and semi-spontaneous speech (SS). From top-left to top-right (following the v-shape): /iː, ɪ, ɛ, a, o, u/.

Table 7: Significant shifts in formant values of vowels in reading task and semi-spontaneous speech (SE = standard error, DF = degrees of freedom).

Vowel	Task	Speaker	Formant	Sessions	Estimate (Hz)	SE (Hz)	DF	p-value
/a/	RT	LS3	F1	A1 – A0	78.8	20.9	110	0.0008
				A2 – A1	–61.8	17.7	110	0.0020
		LS5	F1	A2 – A0	111.1	30.9	107	0.0014
				A2 – A1	84.5	27.4	107	0.0073
		LS1	F2	A1 – A0	–165.3	39.5	110	0.0002
				A2 – A0	–134.5	37.3	110	0.0014
		LS4	F2	A1 – A0	–164.1	44.6	103	0.0011
				A2 – A0	–144.9	42.0	103	0.0024
	SS	LS3	F1	A2 – A1	–44.4	15.3	269	0.0115
		LS4	F1	A1 – A0	–45.9	14.8	215	0.0064
				A2 – A0	–61.3	15.7	215	0.0004
/aː/	RT	LS1	F1	A2 – A0	–62.6	17.2	27	0.0031
				A2 – A1	–65.7	17.2	27	0.0020
		LS3	F1	A2 – A1	–93.7	27.9	25	0.0069
/ɛ/	RT	LS5	F1	A2 – A1	65.6	19.8	152	0.0033
	SS	LS1	F1	A2 – A0	–43.6	12.7	324	0.0019
		LS4	F1	A1 – A0	–56.5	10.8	381	<0.0001
		LS1	F2	A2 – A0	117.5	26.6	324	<0.0001
/ɪ/	RT	LS5	F1	A2 – A0	70.7	20.9	80	0.0031
				A2 – A1	102.0	18.1	80	<0.0001
	SS	LS4	F1	A2 – A0	–26.0	8.56	170	0.0079
		LS3	F2	A2 – A0	–143.4	44.8	222	0.0045
/o/	RT	LS4	F1	A1 – A0	–40.6	13.7	104	0.0103
	SS	LS4	F1	A2 – A1	78.2	26.4	231	0.0094
			F2	A2 – A0	186.6	36.9	231	<0.0001
		LS5	F2	A2 – A0	112.9	37.6	215	0.0083
/u/	RT	LS3	F2	A2 – A0	347.0	75.5	40	0.0001
	SS	LS3	F2	A1 – A0	297.9	62.4	76	<0.0001
				A2 – A0	302.2	62.4	76	<0.0001
		LS5	F2	A2 – A0	217.0	65.3	64	0.0042

3.3 Final schwa

Table 8 shows the number of occurrences of final schwa per speaker and recording session. Speaker LS4 did not use any final schwa at all. Speakers LS1 and LS3 also did

not use any final schwa in the first recording (A0, close to the day of arrival to Toulouse), but later (A2), the final schwa can be found in their speech. Speaker LS5 produced final schwas more often overall, even during the first recording (A0); however, we should note this was recorded 21 days after arriving in Toulouse and in fact, it is actually closer in meaning to A1. Due to the low counts obtained overall, we decided not to conduct a statistical analysis.

Table 8: Number of occurrences of final schwa per speaker and recording session in SS.

	A0	A1	A2
LS1	0	0	1
LS3	0	2	2
LS4	0	0	0
LS5	6	1	6

4. Discussion and conclusions

Our first hypothesis predicted a drift in spectral moments of /ɦ/ and /x/, use of a final schwa in semi-spontaneous speech (SS) and formant shifts of vowels. Table 9 summarises all significant drifts in specific acoustic parameters between two different recording sessions.

We observed the drift in spectral moments of /ɦ/ and /x/ in SS of only one speaker, the obvious drawback is the lack of data to analyse here (see Table 6). The significant drift of the standard deviation of /ɦ/ (between the first and the second recording session, A1 – A0) cannot be judged as assimilation nor dissimilation because /ɦ/ does not exist in French. However, a similar increase of the standard deviation towards the Czech /x/ values in L1 production of late Czech-French bilinguals was found in Hévrová (2021). The significant drift in the skewness of /x/ (between the second and the third recording session, A2 – A1) can be considered as a return back to its original value (A0). Although not statistical significant, the decrease in skewness of /x/between A1 – A0 is also in conformity with findings in Hévrová (2021). We may summarise these observations into two hypotheses: (i) when the phonetic drift/attrition occurs in /ɦ/ and /x/ of Czechs in France, /ɦ/ may be directed towards the spectral moments of Czech /x/, and /x/ may move away from the values of spectral moments of Czech /x/; (ii) the phonetic drift is not linear in time, i.e., some characteristics may evolve in one direction over time with varying speed, and others may also revert later.

Concerning the vowels, we found the drift in F1 of /aː/ as predicted but only for two speakers in RT. However, we also found a drift in F1 or F2 of /a/ for all speakers in RT and an F1 drift for two speakers in SS. We can presume that the lower number of /aː/ items compared to /a/ could influence the lower significant findings in the case of /aː/. We predicted the drift in F1 and F2 of /ɛ/, and we found F1 drift for one speaker in RT and with two other speakers in SS where one of them also had a drift in F2. We also predicted the

drift in F1 and F2 of /ɛː/, but this was not analysed concerning a few /ɛː/ occurrences in both tasks. The predicted drift in F1 and F2 of /ɪ/ was found significant also only in a few cases. In addition, we found some drifts in /o/ and /u/ vowels.

Hence, we may conclude that our first hypothesis was only partially confirmed: the drift appeared in the predicted phonemes but not for all speakers in all phonemes.

The second hypothesis predicted variance in type and amount of the drift among speakers. We summarised all significant drifts in Table 9 for this reason.

We classified the drift of vowels in Table 9 concerning the reference values of Czech and French vowels found in the literature referred to in the caption of Table 2. Assimilation stands for getting closer to the French vowels' values, dissimilation means moving away from the French vowels, and by returning back (not considered as a drift), we mean a movement towards the original value of A0. Although the variety of significant findings across speakers in Table 9 is apparent, we can observe a joint behaviour of drift trends in several cases. In RT, /a/ dissimilates in its F1 (two speakers) and its F2 (two other speakers) while it assimilates in its F1 in SS (two speakers). The /ɛ/ of two speakers assimilates in SS (in F1, or in both F1 and F2), and the /o/ and /u/ dissimilate in their F2 in SS (two speakers). However, as Table 9 shows, the vowels of the speakers often did not drift at the same time: e.g., in RT, the /a/ dissimilates in its F1 between the first and the second recording session of LS3 while in the case of LS5, the similar behaviour is between the second and the third recording session. In SS, it assimilates between the second and the third recording session of LS3 and between the first and the second recording session of LS4. This observation seems to support our second hypothesis of the inter-speaker variance, mainly the assumption of SLM-r that many factors varying across speakers (that means not only the time of L2 immersion) influence together L2 sound production and consequently, the time in which the interactions between L1 and L2 phonetic categories start to occur may vary from speaker to speaker.

Table 9: Significant drifts. Note: A1 – A0 = light grey, A2 – A1 = grey, A2 – A0 = dark grey, N = not analysed, Nd = not determined, A = assimilation, D = dissimilation, B = return back.

	/ɦ/	/x/	/a/				/aː/		/ɛ/		/ɪ/			/o/		/u/	
	SD	skew.	F1		F2		F1		F1	F2	F1		F2	F1	F2	F2	
Reading task (RT)																	
LS1	N				D	D	A	A									
LS3	N		D	B			B									D	
LS4	N				D	D								A			
LS5	N		D	D					B		D	D					
Semi-spontaneous speech (SS)																	
LS1									A	A							
LS3	N		A										D			D	D
LS4	Nd	B	A	A					A		A			B	D		
LS5															D	D	

Acknowledgement

The present study was supported by Grant Agency of Charles University: *Grantová agentura Univerzity Karlovy (GAUK)* – 2020.

REFERENCES

Aoyama, K. & Guion, S. (2007). Prosody in second language acquisition: An acoustic analysis on duration and F0 range. In O.-S. Bohn & M. J. Munro (Eds.), *The role of Language experience in second-Language speech learning: In honor of James Emil Flege* (p. 281–97). Amsterdam: John Benjamins.

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015). Fitting linear mixed-effects models using LME4. *Journal of Statistical Software*, *67*(1).

Bergmann, C., Nota, A., Sprenger, S. A. & Schmid, M. S. (2016). L2 immersion causes nonnative-like L1 pronunciation in German attriters. *Journal of Phonetics*, *58*, 71–86.

Boersma, P. & Weenink, D. (2019). *Praat: doing Phonetics by computer* [Version 6.1.08, retrieved 5 December 2019].

Bořil, T. & Skarnitzl, R. (2016). Tools rPraat and mPraat. In P. Sojka, A. Horák, I. Kopeček & K. Pala (Eds.), *Text, speech, and dialogue: 19th international conference, TSD 2016, Brno, Czech Republic, September 12-16, 2016, Proceedings* (pp. 367–374). Cham: Springer International Publishing.

Brun, A. (2000). *Le français de marseille: étude de parler régional*. Marseille: J. Laffitte.

Candea, M. (2000). *Contribution à l'étude des pauses silencieuses et des phénomènes dits d'hésitation en français oral spontané: étude sur un corpus de textes en classe de français* (PhD Thesis). Université de la Sorbonne nouvelle, Paris.

Čapek, K. (1939). *Měl jsem psa a kočku*. Melantrich.

Čapek, K. (1960). *Jak se co dělá. O lidech*. Prague: Československý spisovatel.

Carton, F., Rossi, M., Autesserre, D. & Léon, P. (1983). *Les accents des francais*. Paris: Hachette.

Chang, C. B. (2010). *First Language Phonetic Drift During Second Language Acquisition* (PhD thesis). University of California, Berkeley.

Chang, C. B. (2012). Rapid and multifaceted effects of second-Language learning on first Language speech production. *Journal of Phonetics*, *40*, 249–268.

Chang, C. B. (2013). A novelty effect in phonetic drift of the native language. *Journal of Phonetics*, *41*, 520–533.

Chang, C. B. (2019). Phonetic drift. In M. S. Schmid & B. Köpke (Eds.), *The Oxford Handbook Language Attrition* (pp. 191–203). Oxford, New York: Oxford University Press.

Colantoni, L. & Steele, J. (2007). Acquiring /r/ in context. *Studies in Second Language Acquisition*, *27*, 381–406.

Coquillon, A. (2005). *Caractérisation prosodique du parler de la région marseillaise* (PhD Thesis). Université Aix-Marseille I, Marseille.

Curtin, S., Goad, H. & Pater, J. (1998). Phonological transfer and levels of representation: the perceptual acquisition of Thai voice and aspiration by English and French speakers. *Second Language Research*, *14*, 389–405.

De Leeuw, E. (2008). *When your native language sounds foreign: A phonetic investigation into first language attrition* (PhD Thesis). Queen Margaret University, Edinburgh.

De Leeuw, E. (2019). Phonetic attrition. In M. S. Schmid & B. Köpke (Eds.), *The Oxford Handbook of Language Attrition* (pp. 204–217). Oxford, New York: Oxford University Press.

De Leeuw, E., Schmid, M. S. & Mennen, I. (2010). The effects of contact on native language pronunciation in an L2 migrant setting. *Bilingualism: Language and Cognition*, *13*, 33–40.

De Leeuw, E., Tusha, A. & Schmid, M. S. (2018). Individual phonological attrition in Albanian–English late bilinguals. In: Bilingualism: Language and Cognition (Vol. 21, p. 278–295). Cambridge University Press.

Durand, J. (2009). Essai de panorama phonologique: Les accents du midi. In L. Baronian & F. Martineau (Eds.), *Le français d'un continent à l'autre. mélanges offerts à yves-charles morin* (pp. 123–170). Québec: Presse de l'Université Laval.

Durand, J., Slater, C. & Wise, H. (1987). Observations on schwa in Southern French. *Linguistics*, *25*(2), 983–1004.
Elvin, J., & Escudero, P. (2019). Cross-Linguistic Influence in Second Language Speech: Implications for Learning and Teaching. In G.-M. M., M.-A. M., & G. del Puerto F. (Eds.), *Cross-Linguistic Influence: From Empirical Evidence to Classroom Practice. Second Language Learning and Teaching* (pp. 1–20). Springer International Publishing, Cham.
Escudero, P. (2005). Linguistic perception and second Language acquisition (PhD Thesis). Utrecht University, Utrecht.
Flege, J. E. (1987). The production of "new" and "similar" phones in a foreign language: Evidence for the effect of equivalence classification. *Journal of Phonetics*, *15*, 47–65.
Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (p. 233–277). Timonium, MD: York Press.
Flege, J. E. & Bohn, O.-S. (2021). The revised speech learning model (SLM-r). In R. E. Wayland (Ed.), *Second Language speech learning: Theoretical and empirical progress* (p. 3–83). Cambridge: Cambridge University Press.
Gallo, F., Bermudez-Margaretto, B., Shtyrov, Y., Abutalebi, J., Kreiner, H., Chitaya, T., Petrova, A., Myachykov, A. (2021). First language attrition: What it is, what it isn't, and what it can be. *Frontiers in Human Neuroscience*, *15*, 686388.
Gendrot, C. & Adda-Decker, M. (2005). Impact of duration on F1/F2 formant values of oral vowels: An automatic analysis of large broadcast news corpora in French and German. In *Interspeech 2005* (p. 2453–2456). Lisbon, Portugal.
Herd, W., Walden, R., Knight, W. & Alexander, S. (2015). Phonetic drift in a first language dominant environment. *Proceedings of meetings on acoustics Acoustical Society of America*, *23*.
Hévrová, M. (2021). *Phonetic attrition and cross-linguistic influence in L1 speech of late Czech-French bilinguals* (PhD Thesis). Université Toulouse 2 – Jean-Jaurès – Charles University, Toulouse-Prague.
Hévrová, M., Bořil, T. & Köpke, B. (2020). Phonetic attrition in vowels' quality in L1 speech of late Czech-French Bilinguals. In P. Sojka, I. Kopeček, K. Pala & A. Horák (Eds.), *Text, speech, and dialogue. TSD 2020. Lecture notes in computer science, vol. 12284* (pp. 348–355). Cham: Springer International Publishing.
Holliday, J. J. (2015). A longitudinal study of the second language acquisition of a three-way stop contrast. *Journal of Phonetics*, *50*, 1–14.
Jarvis, S. & Pavlenko, A. (2008). *Crosslinguistic influence in language and cognition*. New York: Routledge.
Kartushina, N., Hervais-Adelman, A., Frauenfelder, U. & Golestani, N. (2016). Mutual influences between native and non-native vowels in production: Evidence from short-term visual articulatory feedback training. *Journal of Phonetics*, *57*, 21–39.
Kijak, A. (2009). *How stressful is L2 stress? A cross-linguistic study of L2 perception and production of metrical systems*. Utrecht: Landelijke Onderzoekschool Taalwetenschap.
Köpke, B. (2004). Attrition is not a unitary phenomenon. On different possible outcomes of language contact situations. In F. R. Anxo M. L. Suarez & X. P. Rodiguez-Yanez (Eds.), *Bilingual socialization and bilingual language acquisition. Proceedings from the second international symposium on bilingualism* (p. 1331–1347). Vigo: Universidade de Vigo.
Köpke, B. & Schmid, M. S. (2004). Language attrition: The next phase. In M. S. Schmid, B. Köpke, M. Keijzer & L. Weilemar (Eds.), *First Language attrition. Interdisciplinary perspectives on methodological issues* (pp. 1–43). Amsterdam: John Benjamins.
Kornder, L. & Mennen, I. (2021). Longitudinal developments in bilingual second language acquisition and first language attrition of speech: The case of Arnold Schwarzenegger. *Languages*, *6*(2), 1–25.
Kupske, F. F. & Alves, U. K. (2016). A fala de imigrantes brasileiros de primeira geração em londres como evidência empírica para a língua como um sistema adaptativo complexo. *Revista virtual de estudos da linguagem*, *14*(27), 173–203.
Lang, B. & Davidson, L. (2019). Effects of exposure and vowel space distribution on phonetic drift: Evidence from American English learners of French. *Language and Speech*, *62*(1), 30–60.

Lenth, R. V. (2021). *emmeans: Estimated marginal means, aka least-squares means.* (R package version 1.5.4)
Léon, M. (1997). *La prononciation du français*. Paris: Nathan.
Machač, P. & Skarnitzl, R. (2009). Fonetická segmentace hlásek. Praha: Epocha.
Major, R. C. (1986). The ontogeny model: Evidence from L2 acquisition of Spanish r. *Language Learning*, *36*(4), 453–504.
Major, R. C. (1992). Losing English as a first language. *The Modern Language Journal*, *76*(2), 190–208.
Mayr, R., Price, S. & Mennen, I. (2012). First language attrition in the speech of Dutch–English bilinguals: The case of monozygotic twin sisters. *Bilingualism: Language and Cognition*, *15*(4), 687–700.
Mayr, R., Sánchez, D. & Mennen, I. (2020). Does teaching your native language abroad increase L1 attrition of speech? The case of Spaniards in the United Kingdom. *Languages*, *5*(4), 1–41.
Paillereau, N. (2015). *Perception et production des voyelles orales du français par des futures enseignantes tchèques de Français Langue Etrangère (FLE)* (PhD Thesis). Université Sorbonne-Nouvelle, Paris.
Paillereau, N. & Chládková, K. (2019). Spectral and temporal characteristics of Czech vowels in spontaneous speech. *AUC PHILOLOGICA*, *2019*(2), 77–95.
Pavlenko, A. (2000). L2 influence on L1 in late bilingualism. *Issues in Applied Linguistics*, *11*(2), 175–205.
Průchová, T. (2016). *Charakteristika hezitací v češtině a míra jejich percepční rušivosti* (Bachelor thesis). Charles University, Faculty of Arts, Prague.
Pustka, E. (2011). L'accent méridional: représentations, attitudes et perceptions toulousaines et parisiennes. *Lengas. Revue de sociolinguistique*(69), 117–152.
R Core Team. (2019). *R: A Language and environment for statistical computing.* Vienna, Austria.
Rubovičová, C. (2014). *Tempo řeči a realizace pauz při konsekutivním tlumočení do češtiny ve srovnání s původními českými projevy* (Master thesis). Charles University, Faculty of Arts, Prague.
Sancier, M. L. & Fowler, C. A. (1997). Gestural drift in a bilingual speaker of Brazilian Portuguese and English. *Journal of Phonetics*, *25*(4), 421–436.
Sedláčková, P. (2010). *Srovnání vybraných spektrálních a temporálních vlastností německých frikativ [x] a [ç] a české frikativy [x] a jejich odezva v percepci* (Master thesis). Charles University, Faculty of Arts, Prague.
Sharwood Smith, M. (1983). Cross-linguistic aspects of second language acquisition. *Applied Linguistics*, *4*(3), 192–199.
Skarnitzl, R. & Volín, J. (2012). Referenční hodnoty vokalických formantů pro mladé dospělé mluvčí standardní češtiny. *Akustické listy*, *18*, 7–11.
Skarnitzl, R., Šturm, P. & Volín, J. (2016). *Zvuková báze řečové komunikace: Fonetický a fonologický popis řeči*. Prague: Karolinum.
Sůčková, M. (2020). *Phonetic and lexical features of first language attrition: Research into the speech of anglophone expatriates in the Czech republic* (PhD Thesis). Masaryk University Faculty of Arts, Brno.
Stoehr, A., Benders, T., van Hell, J. G. & Fikkert, P. (2017). Second language attainment and first language attrition: The case of VOT in immersed Dutch–German late bilinguals. *Second Language Research*, *33*(4), 483–518.
Šulecová, D. (2015). *K parazitickým slovům v české slovní zásobě* (Master thesis). University of South Bohemia in České Budějovice, Faculty of Education, České Budějovice.
Tobin, S., Nam, H. & Fowler, C. (2017). Phonetic drift in Spanish-English bilinguals: Experiment and a self-organizing model. *Journal of Phonetics*, *65*, 45–59.
Tubach, J. P. (1989). *La parole et son traitement automatique*. Paris, Milan, Barcelone: Masson.
Tykalová, T., Škrabal, D., Bořil, T., Čmejla, R., Volín, J. & Rusz, J. (2021). Effect of Ageing on Acoustic Characteristics of Voice Pitch and Formants in Czech Vowels. *Journal of Voice, 35*(6), 931.e21-931.e33.
Ulbrich, C. & Ordin, M. (2014). Can L2-English influence L1-German? The case of post-vocalic /r/. *Journal of Phonetics*, *45*, 26–42.
Van Leussen, J.-W., & Escudero, P. (2015). Learning to perceive and recognize a second language: the L2LP model revised. Frontiers in Psychology, 6, 1000.
Volín, J. (2010). Fonetika a fonologie. In V. Cvrček (Ed.), *Mluvnice současné češtiny* (p. 35–64). Prague: Karolinum.
Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag New York.
Wickham, H., François, R., Henry, L. & Müller, K. (2020). *Dplyr: A grammar of data manipulation.* (R package version 0.8.4)

Woehrling, C. (2009). *Accents régionaux en français: perception, analyse et modélisation à partir de grands corpus* (PhD Thesis). Université Paris Sud – Paris XI, Paris.

Marie Hévrová
Institute of Phonetics
Faculty of Arts, Charles University
Prague, Czech Republic
E-mail: mariehevrova@gmail.com

Tomáš Bořil
Institute of Phonetics
Faculty of Arts, Charles University
Prague, Czech Republic
E-mail: tomas.boril@ff.cuni.cz

ACTA CAROLINAE PRAGENSIS
PHILOLOGICA 1/2022

Editors: Jan Volín
Pavel Šturm
Cover and layout by Kateřina Řezáčová
Published by Charles University
Karolinum press, Ovocný trh 560/5, 116 36 Praha 1
www.karolinum.cz
Prague 2022
Typeset by Karolinum Press
Printed by Karolinum Press

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