ACTA UNIVERSITATIS CAROLINAE
PHILOLOGICA 3/2025

ACTA UNIVERSITATIS CAROLINAE

PHILOLOGICA 3/2025

RADEK SKARNITZL
JAN VOLÍN (eds.)

CHARLES UNIVERSITY
KAROLINUM PRESS

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

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

CONTENTS

EDITORIAL

After three years, AUC Philologica returns to the immense research area of sound patterns in speech. The contents of the current issue have been coordinated by the Institute of Phonetics in Prague with invitations to contributors from other institutions and other countries. Importantly, all the reviewers were phoneticians from outside not only the Institute, but also outside Charles University. We would like to thank all these external scholars for their selfless hard work. Their expertise inspired improvement on many aspects of the submissions, but also prevented publication of a text that they considered of a low standard.

The resulting eleven articles provide a variety of topics, methods and scientific approaches to sound forms in speech. In terms of their focus, they could be divided as follows: a triplet representing applied phonetics, a triplet dedicated to the Czech language, a pair of papers dealing with various issues of second/foreign language acquisition, a pair of papers concerning languages other than Czech, and one contribution from the field of history of phonetics.

The 'applied' triplet starts with Michaela Svatošová and Jan Volín's contribution to the problem of local articulation rate, which they believe reflects an important part of prosodic structure. They offer a description of a tool that will allow for the visualization of articulation rate courses and for more informed measurements of local events. Jan Šebek and Tomáš Bořil present a study of voice changes in cancer patients who underwent radiotherapy. Their longitudinal research combines acoustic measurements with perceptual tests and provides facts that could lead to improved post-therapeutic treatment of patients in the future. Finally, Adléta Hanžlová and Václav Hanžl present their work on an automatic speech sound aligner for Czech spoken texts. The results generated by this extremely useful tool for phonetic research are undoubtedly encouraging.

The triplet of papers that deal with the Czech language starts with Pavel Šturm's account of temporal properties of repeated parts in spoken texts. He used performances of speakers reciting poetry to establish temporal changes in repetitions. Since his findings are related to the information structure of utterances, we arranged Jan Volín's probe into stress patterns in modal verbs as the following article. He, too, relates his research to the information value of utterance constituents. Modal verbs are neither typical auxiliaries nor autosemantic verbs, so it is of interest to know how often their stress potential materializes in continuous narratives spoken by professional actors. The subject matter of the third paper on Czech is speech tempo. The study submitted by Jitka Veroňková examines

performances of students in news reading and semi-spontaneous monologues, and confronts the findings with results from previous research.

The next pair of papers tackles issues in second or foreign language acquisition. The team led by Alice Henderson presents an analysis of comments by learners of English who come from Central and South America. The objective was to map students' attitudes to the concepts of fluency and intelligibility in order to inform didactic efforts of teachers of English. Anna Chabrová, on the other hand, focuses on German learners of Czech and examines the interference of the German vowel system in the Czech spoken as a foreign language.

In the pair of papers on languages other than Czech, Nela Bradíková and Radek Skarnitzl investigated American and British English in the genre of political debate. Their focus was the behaviour of stressed and unstressed vowels. Using highly attractive real-life recordings, they confront their findings with theories stemming from laboratory experiments. Another language thematized in the current issue of AUC Philologica is Dogri, spoken in Northern India: Pranav Badyal examines singleton and geminate consonants to see how their phonological opposition manifests in speech.

Finally, there is a remarkable contribution by Pavel Šturm and Jürgen Trouvain, who devote their efforts to the history of phonetics. Their study is based on interviews of two senior phoneticians reflecting their academic careers. It delivers inspiring reading to anyone who feels to belong to the field of phonetics.

We wish all the readers of this issue pleasant and enriching moments with the articles here.

Jan Volín and Radek Skarnitzl
doi: 10.14712/24646830.2025.17

ARTICLES

2025 ACTA UNIVERSITATIS CAROLINAE PAG. 11–31
PHILOLOGICA 3

EXPLORING THE PHRASE-INTERNAL CHANGES IN ARTICULATION RATE: THE LAROMETER TOOL AND ITS APPLICATIONS

MICHAELA SVATOŠOVÁ, JAN VOLÍN

ABSTRACT
The article introduces a method of normalising the inherent durational properties of phones (the LARometer), providing a relative measure of *local articulation rate* (LAR). It allows for the quantification of the communicatively relevant variations in articulation rate and their visualisation in temporal contours. The normalisation is based on an extensive manually annotated corpus containing over four hours of continuous speech. The usage of the LARometer is illustrated with two studies. Study 1 identifies locally decelerated content words in Czech radio news reading. These decelerated words often had prominent functions in the information structure (rheme, contrastive topic). The results also indicated that deceleration affects various parts of words. Study 2 focuses on phone reductions in television political debates. The reductions were predominantly observed in content words, but function words were affected to greater extent. Also, considerable differences in the number of reductions were found between individual speakers.

Keywords: local articulation rate; temporal variability; phone duration; normalisation; information structure

1. Introduction

Speech prosody entails the study of all domains of sound that evolve in time, including tempo. The rate of articulating speech units in natural speech production is clearly not constant, moreover, many of the factors causing this variation are linked to communicative functions. The affective state of the speaker (being happy, nervous, sad, bored etc.) influences the global tempo (Trouvain, 2003, p. 15). In conversations, higher articulation rate is linked to parenthetical structures, i.e. utterances containing less important information (Local, 1992; Uhmann, 1992). The temporal cues also contribute to phrasing through the insertion of pauses and phrase-final deceleration (lengthening). Additionally, experiments with elicited sentences have shown that words in focus are articulated more slowly than in other information structure roles (Baumann et al., 2007; Cooper et al., 1985; Heldner & Strangert, 2001). Further research on the functional uses of the

https://doi.org/10.14712/24646830.2025.18

temporal variation therefore requires a reliable measure for assessing these changes, both global and local (phrase-internal).

The notion of *tempo* or *rate*[1] is underspecified and refers to a range of distinct subtypes. Depending on the treatment of pauses, speech rate (including pauses) or articulation rate (excluding pauses) can be distinguished. Both can be understood as realised or canonical, taking into account only the articulated segments, or the standard form of the uttered word. In either case, tempo is usually quantified as the number of speech units (words, syllables, phones) per a given time frame (minute, second). The measures typically characterise a longer stretch of speech with one mean value, e.g. the mean articulation rate of 6 syllables per second in the whole text produced by a given speaker. However, such averaging conceals and underestimates the variability of articulation rate (Miller et al., 1984) and the resulting values might be insufficient for capturing important, but more complex patterns (cf. the explanatory value of a melodic rise vs. its mean F0).

Unfortunately, calculating means of syllables or phones per second on a local level (in short phrases or individual words) is distorted by the specific composition of the given extract. Phone durations are affected by their inherent characteristics including vowel length and height, obstruent voicing or manner of articulation of consonants. These factors cancel each other out in longer stretches of speech, which contain phones of all kinds, but they can have a strong effect in short stretches, making comparisons of mean articulation rates between different words problematic.

Since comparing articulation rates of different words is inevitable and necessary especially in studies on spontaneous speech, new approaches to quantifying articulation rate have emerged. Saarni et al. (2008) presented a relative measure, which groups phones into seven classes and normalises the duration of each phone to the mean duration of phones belonging to its class. The mean durations of each class were calculated based on phones in the interpausal unit that was being examined, resulting in potentially unreliable values due to a relatively small number of observations in each interpausal unit. A slightly different approach was proposed by Campbell (2000, pp. 310–311), who transformed the phone durations to *z*-scores, thus expressing them as the number of standard deviations from the mean. The means and standard deviations of all English phones were obtained in advance from an annotated speech corpus, providing statistically more reliable values that could be used as a common reference for normalising the articulation rate of any English utterance.

This article introduces the LARometer tool (LAR = Local Articulation Rate), which is based on similar principles as Campbell's approach. It normalises inherent durational characteristics of Czech phones and provides a measure for capturing local changes in articulation rate. The use of this method is then illustrated with two example studies that explore the relationship between the local articulation rate and information structure (Study 1) and phone reductions in spontaneous speech (Study 2).

[1] In order to differentiate between the objective and subjective aspects of speech, *rate* is used for the objectively measurable characteristics of articulation and *tempo* for their perceptual impact on the listener throughout the article (in analogy to the difference between F0 contours and intonation).

2. LARometer: a model for calculating the local articulation rate

As indicated above, the main aim of the proposed metric is to allow for evaluating changes in articulation rate on a local level, within individual phrases (although its uses do not need to be limited to that). More specifically, the LARometer should represent a tool for describing the prosodically relevant local changes in articulation rate. Prosodic relevance here refers to the fact that unlike simple rate metrics, the LAR values are normalised for some of the durational characteristics of phones that are inherently present in speech and thus cannot express any intentional information from the speaker. Since the LARometer produces quantitative values, these can be used in statistical analyses or visualised in the form of rate contours.

2.1 Material

Durational characteristics of phones are based on physiological factors (including vowel height or obstruent voicing), but they might also exhibit language-specific features. Prior to creating the model, we therefore collected material that would be extensive enough to provide a sufficient amount of durational data to describe the inventory of Czech phones. To make the results ecologically valid, we aimed at continuous speech with a communicative intent, since durational characteristics of words and phrases produced in isolation significantly differ from those of continuous utterances (Klatt, 1975, p. 138; Wagner et al., 2015).

We chose recordings of two genres – radio news (NWS) and storytelling in audiobooks (STR). Each genre was represented by 16 speakers. The radio news were short accounts of current affairs that were broadcast by professional news presenters on the national stations Czech Radio I and II. Two to three (complete) news bulletins were used for each speaker. The audiobooks were produced by professional actors and included both texts written by Czech authors and literature translated to Czech from other languages. A continuous excerpt containing at least 1,000 words (corresponding to approximately 5,500 phones on average) was extracted for each speaker. Further characteristics regarding the size of the material are provided in Table 1.

Table 1 Overview of the material used to obtain the durational characteristics of Czech phones, which consisted of radio news (NWS) and storytelling in audiobooks (STR).

Genre	Speakers	Words	Phones (all)	Phones (analysed)	Speech time (minutes)
NWS	16 (8 M, 8 F)	16,778	97,730	83,785	118
STR	16 (8 M, 8 F)	16,939	78,824	62,956	134
Total	32 (16 M, 16 F)	33,717	176,554	146,741	252

In total, the material contained over 170,000 phones. The placement of their boundaries was determined automatically at first (Pollák et al., 2007) and then manually corrected in Praat (Boersma & Weenink, 2024) according to the guidelines summarised by

Machač and Skarnitzl (2009) to ensure highly precise results, because durational measurements form the basis of the presented method.[2]

The material was subsequently restricted by excluding phones in phrase-final positions (from the nucleus of the penultimate syllable to the end of the phrase), which tend to be affected by final lengthening, together with plosives and affricates following a pause, whose duration cannot be determined from a spectrogram. The durational measurements were thus based on the remaining 146,741 phones. The data were processed in R using the packages *rPraat*, *tidyverse* and *patchwork* (R Core Team, 2024; Bořil & Skarnitzl, 2016; Wickham et al., 2019; Thomas Lin Pedersen, 2024).

2.2 Component 1: Inherent duration of phones

In order to obtain the durational characteristics for the inventory of Czech phones, the data were summarised in two steps. Firstly, we calculated the mean duration of a given phone for each speaker in the material. Secondly, these speaker-specific means were averaged to produce a final grand mean value associated with that phone. This procedure was applied to each phone separately. Log-transformed values of duration were used in all calculations, since their distribution more closely resembled the normal distribution.

Individual phones have significantly disparate frequencies of occurrence in spoken texts, which led to uneven numbers of their realisations in the material. For the 24 most common phones, the grand means were based on data from all speakers with at least 30 realisations per speaker. However, these criteria could not be maintained in the case of the less common phones. The grand means were obtained from 24 or more speakers (with at least 10 realisations per speaker) for 14 such phones, e.g. [ɛː uː ʒ f g]. Finally, only 5–28 speakers could provide 3–30 realisations for the rare phones [a͡u oː l̩ ŋ d͡z d͡ʒ ɣ]. The Czech inventory also includes the diphthong [ɛ͡u] and the sonorants [m̩ ɱ]. Due to the lack of data, the LARometer handles these by analogy with [o͡u] and [m]. The grand mean values for the less common phones might not be as reliable as for the others, but their verification would require a larger corpus of comparable speech material. Nevertheless, the potential imperfections should not distort the performance of the LARometer greatly, because the frequency of occurrence of these phones in texts is extremely low.

The observed inherent durations (grand means) complied with expectations based on previous research. Phonologically long vowels had longer duration than short vowels and they also showed an effect of vocalic height, with the low [aː] being the longest and the high [iː uː] being the shortest among the long vowels. Syllabic liquids were longer than non-syllabic ones and voiced obstruents were shorter than their voiceless counterparts. Inherent durations differed also with respect to the place of articulation.

---

[2] Future studies might explore whether manually checked boundaries provide better input for the LARometer or whether the tool is robust enough to work with automatically segmented data. Nevertheless, the durational values included directly in the LARometer (described in the following section) should be as reliable as possible.

2.3 Component 2: Inherent duration of phone bigrams

In addition to its identity, the duration of a phone in continuous speech is influenced by its phonetic context – consonants tend to be shorter in clusters than intervocalically, vowels are often longer in open syllables than followed by a coda. In accordance with the aim of describing only the prosodically relevant changes in articulation rate, the LARometer should normalise this variation as well, since it is automatically related to the segmental composition of utterances regardless of the speaker's intentions. To achieve this, the LARometer was expanded with a second component, which addressed phone bigrams.

The process of measuring inherent durations described in the previous section was therefore applied to pairs of neighbouring phones (bigrams), considering each pair as a single unit. However, obtaining bigrams in sufficient numbers of occurrences in natural texts poses an even greater challenge than with individual phones, due to the much wider range of possible phone combinations. Many underrepresented bigrams were thus not included into the LARometer's second component. In our material, there were 173 bigrams produced by 10 or more speakers (with at least 10 realisations per speaker). These accounted for 61% of the realised bigrams in the utterances, although they represented only 13% of all unique bigrams found in the material (a result of the huge differences in their text frequency).

It has been often noted that the placement of phone boundaries is rather arbitrary, especially for phone combinations without abrupt spectral changes (e.g., approximants adjacent to vowels). In such cases, considering the duration of the whole bigram instead of the individual phones is certainly more convenient and perhaps also more appropriate. Sequences of the sonorants [j ɲ] preceded or followed by any of the vowels [ɪ ɛ iː ɛː] were identified as the most difficult to segment and they were included in the list of bigrams regardless of their frequency in the material. In practice, there were only 6 bigrams which did not meet the aforementioned criteria, so the second component was based on inherent durations of 179 bigrams.

The measurements supported the statements mentioned earlier – for example, the inherent durations of the bigrams [st sk kt pr̥ mɲ] were 10–15% shorter than the simple sum of the respective inherent durations of these phones. It could be argued that a more adequate approach would be to consider the phone's position in the syllable instead of combining it with the adjacent phones, since the present method did not differentiate between an onset–onset sequence and a coda–onset sequence (for pairs of consonants, but similarly with all phones). However, Czech has a very complex syllabic structure (typically (CC)V(C), but more consonants in onset or coda are possible), which yields over 16,000 attested syllables (Šturm & Bičan, 2021, pp. 330–331) and makes automatic assignment of syllabic boundaries unreliable. Adhering to simple bigrams does not require the additional layer of annotation for syllables, keeping the method more widely usable.

2.4 Calculation of the local articulation rate (LAR)

This section introduces the process of calculating the local articulation rate values and contrasts it with the common articulation rate measures. For illustration, Figure 1 pre-

sents rate contours of one prosodic phrase (chosen from the radio news material) based on the described metrics. The exact values used in the calculation of the LAR values for a single word from that phrase are shown in Table 2. For the sake of clarity, the following tables and figures display durations in milliseconds, although the LARometer internally uses log-transformed values.

The simple non-normalised articulation rate metrics divide the number of phones in a given interval by the duration of that interval. For the example phrase in Figure 1, this leads to the overall articulation rate of 14.3 phones/second (indicated as the grey dotted line in Panel A). In order to achieve a more local perspective, however, the size of the interval needs to be diminished. Mean values can be calculated for individual words (dashed horizontal lines). In the extreme case, only a single phone would be considered at a time, which would then correspond to the formula in (1). Since this approach does not normalise for the inherent durations of phones, it results in seemingly fast articulation rates for short phones like [v j n ɪ] and slow articulation rates for long phones like [t͡ʃ s]. Comparing articulation rate means of different words is also problematic, since they are affected by the number and type of phones they contain.

(1) $AR_{phone} = \frac{1}{dur_{phone}}$

The LARometer therefore computes the local articulation rate values differently, using the formula in (2). It relates the observed duration of each unit in question (dur_{obs}) to its inherent duration (dur_{inh}), which was estimated on the basis of a large corpus of speech (as described in the previous sections).[3]

(2) $LAR = \frac{dur_{inh}}{dur_{obs}}$

Table 2 The durational values used for the calculation of the local articulation rate (LAR) values of the word '*končí*' from the example phrase in Figure 1. See Formula (2) and the text for details.

Phone	Observed duration (ms)	Inherent duration (ms)		Local articulation rate
k	72	136		0.98
o	67		109	1.03
n	33			1.09
t͡ʃ	110	114		1.04
iː	80	68		0.85

The calculation of the LAR value for each phone is primarily based on the durations of the two bigrams of which it is a part – e.g., the final value for [o] in the word '*končí*' was obtained as a mean of the LAR values for the bigrams [ko] and [on] (see Table 2). If the list of inherent durations for bigrams includes only one of the two phone sequences

[3] The value of the inherent duration is in the numerator, which yields higher LAR values for faster articulation rate and lower LAR values for slower articulation rate. Since LARometer uses log-transformed durational measures, the formula is in fact $LAR = dur_{inh} - dur_{obs.}$

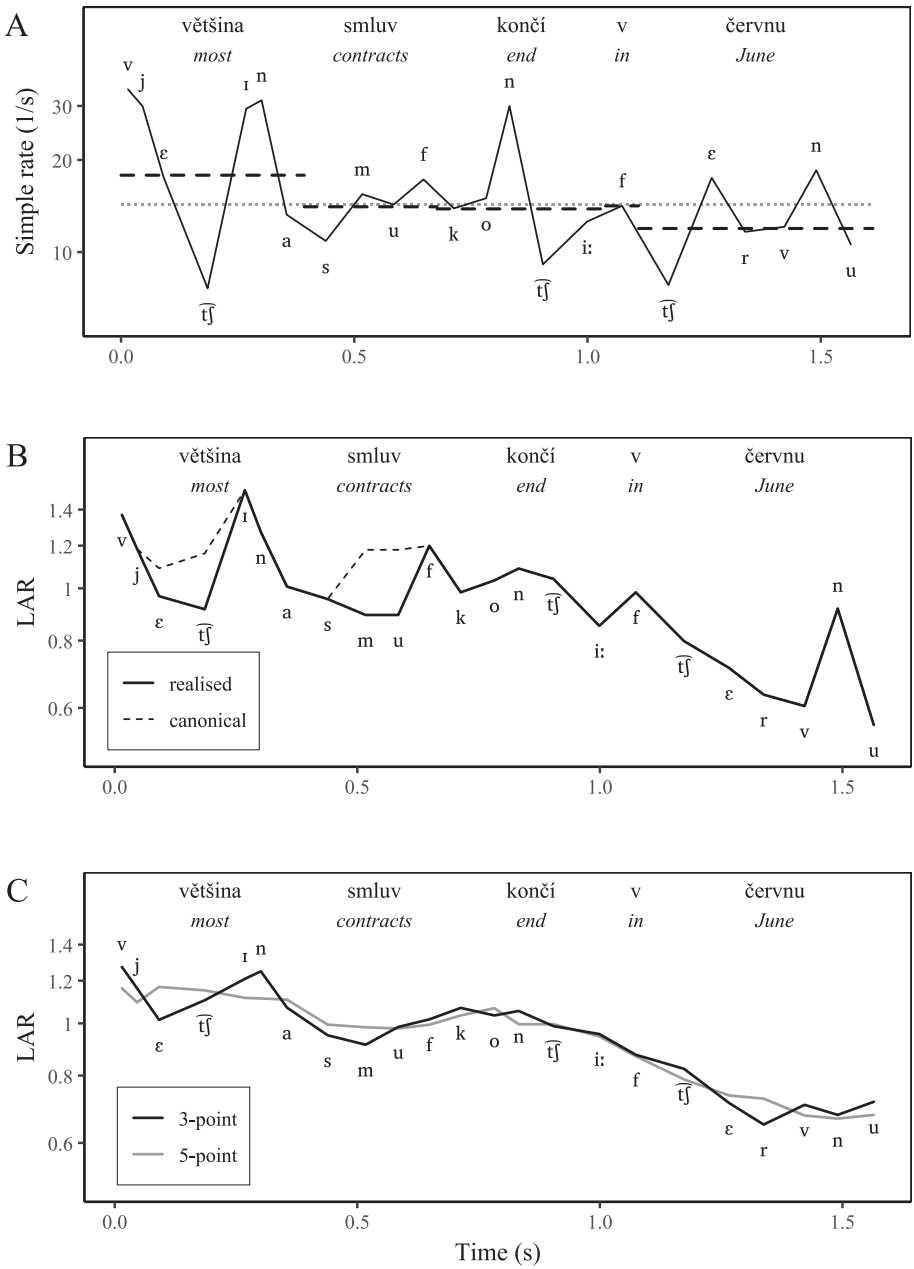


Figure 1 The simple rate and local articulation rate (LAR) contours of an example phrase. Panel A: mean articulation rate of the phrase (dotted line), simple articulation rate in words (dashed line) and by phones (solid line). Panel B: realised (solid line) and canonical (dashed line) local articulation rate contours. Panel C: realised LAR contour smoothed with a 3-point (black line) and a 5-point (grey line) moving average.

in consideration, the LAR value of that bigram is used as the final value for the target phone. This is the case for the phone [n] in the example – since the list of bigrams does not contain the bigram [nt͡ʃ], the final LAR value is based only on the bigram [on]. Sometimes neither of the two bigrams is on the list and the LAR value has to be calculated by relating the observed and inherent duration of the target phone only. This was done for the phone [t͡ʃ], because neither of the sequences [nt͡ʃ] and [t͡ʃi:] is on the list of bigrams.

For convenience and visualisation, the raw LAR values (on a log-scale) can be transformed to ratios or percentages. These are plotted for the whole phrase in Panel B of Figure 1 (solid line). The value 1 (or 100%) expresses that the duration of the given phone was equal to its inherent duration. Smaller LAR values correspond to deceleration (slower articulation rate), e.g. 0.85 indicates that the phone's articulation rate was 85% of the average rate based on inherent durations, whereas phones articulated with a faster rate have higher LAR values. In contrast to the simple rate contour, the LAR contour shows less abrupt changes between neighbouring phones and incorporates the seemingly outlier phones into the contour (compare [n] or [t͡ʃ] in the simple rate and LAR contours). Moreover, an overall trend of deceleration emerges from this picture, which was obscured by the amount of variation present in the simple rate contour.

The calculations described so far reflect the *realised* articulation rate and they require only the phone-level annotation. If phonemes[4] are provided as well, the LARometer can additionally compute the *canonical* local articulation rate (Koreman, 2006; Plug et al., 2022). The procedure remains principally the same, but the observed duration of the realised bigram or phone is related to the expected duration of the phone(s) that would be pronounced canonically. For example, the word '*smluv*' was reduced to [smuf] instead of the canonical [smluf] in the described phrase (elision of [l]). For the realised LAR values of the phones [m] and [u], the observed duration of the sequence [mu] was compared to its inherent duration. However, the canonical LAR was based on relating the observed duration of [mu] to the expected duration of the sequence [mlu] (the combinations [sm] and [uf] are not on the list of bigrams).

The canonical local articulation rate values are represented by the dashed line in Panel B of Figure 1. They differ from the realised LAR in any case of mismatch between the annotation of phones and phonemes, e.g. elisions ([smuf] instead of [smluf] described above), phone alternations (substitution of the sequence [tʃ] with the affricate [t͡ʃ] in the word *'většina'*) or epentheses (not present in the example phrase).

Finally, the realised and canonical LAR values can be further modified, e.g. by smoothing. Panel C of Figure 1 illustrates the contour of realised local articulation rate smoothed with a 3-point (black line) and a 5-point (grey line) moving average. Rate contours smoothed by a 3-point moving average might provide a reasonable compromise between constraining occasional outliers (e.g., the [n] in the final word) and preserving the local changes of LAR. Subsequently, these smoothed values can be averaged in higher-level units (syllables, words, accent-groups, phrases etc., see examples in Study 1 below) and also used in statistical analyses. Importantly, the resulting mean LAR values

[4] Phonemes are considered here canonical constitutive units of lexemes as specified in a standard lexicon.

are not biased by the segmental composition of the respective units, which allows for comparisons of units with different length and content.

The interpretation of the LAR values is in many respects analogous to semitones. Most importantly, the reference value (1 or 100%) bears no definite meaning in itself. It is therefore advisable to find a reference that would be meaningful for the particular research question (e.g., a mean LAR value in each phrase or for each speaker that is being analysed) and to normalise all calculated values to that reference. The key product of the LARometer are the relations between local articulation rates of phones, which remain intact by this procedure. If a given phone A is pronounced twice as fast as another phone B, the same relation can be expressed with the LAR values 1.0 and 0.5 (the reference being the phone A) or with the LAR values 2.0 and 1.0 (the reference being the phone B).[5]

3. Study 1: Locally decelerated content words in radio news reading

3.1 Aims

The LARometer was created with the purpose of describing local changes in articulation rate of prosodic phrases. This section describes an exploratory study that was conducted in order to test its possibilities. Slower articulation rate corresponds to higher prominence, but articulation rate is known to be affected by other factors as well. Locally decelerated (and therefore potentially prominent) content words were identified in a corpus of radio news. The main aim of this study was to analyse them in terms of their phonetic characteristics and role in the information structure and to explore whether these could be meaningfully explained in relation to each other.

3.2 Method

The material used for this study consisted of radio news extracts provided by 16 professional radio presenters (see Section 2.1). This genre contains authentically produced continuous speech with the purpose of providing information to audiences. The speakers have an opportunity to familiarize themselves with the text before the broadcast begins. As a result, they could be expected to use various prosodic cues to deliver the intended meaning as clearly as possible.

There were 3,451 major prosodic phrases in total (ToBI-4 according to Beckman & Ayers Elam, 1993), however, the study used only a subset defined by the following criteria. Phrases consisting of multiple minor phrases (ToBI-3, n = 1,094) were excluded, since the intermediate boundary could be accompanied with a deceleration that was not of interest here. There was also a limit on the number of accent-groups. Comparing relative articulation rates of words makes little sense in very short phrases. On the other hand, there were only two phrases with 9 and 11 accent-groups. As a result, we worked with phrases containing 3 to 8 accent-groups (n = 1,284).

[5] On the logarithmic scale, this relation is captured by the difference of $\log_{10}(2/1)$, which equals cca 0.3.

The recordings were phonetically annotated on the level of words and phones with manual corrections of phone boundaries. The values of the realised local articulation rate (LAR) were computed for the whole dataset using the LARometer algorithm (see Section 2) and they were subsequently smoothed with a 3-point moving average. Furthermore, all values were normalised relative to the mean LAR in the given phrase (excluding the final accent-group), which enabled comparable analyses across phrases and speakers.

The present study aimed to examine locally decelerated content words before the final accent-group, i.e. the slowest content word of the phrase that was not decelerated due to phrase-final lengthening. These target words were identified with two approaches. Firstly, the LAR values were averaged in words and the content word with the slowest mean LAR (in a non-final accent-group) was taken as the target word (the 'WORD' method). Secondly, the slowest phone of the phrase was found (also disregarding the final accent-group) and the word containing that phone was considered the target word (the 'PHONE' method).

The target words were then ordered by their phrase-normalised LAR. In order to allow for a qualitative analysis of information structure, 5% ($n = 64$) of cases were selected from the lists provided by each method. These were words with the slowest mean LAR value in the word (the WORD method) or with the slowest phone (the PHONE method) relative to the mean LAR of the phrase. The selection disregarded abbreviations, which are pronounced in a specific style, and words prolonged due to hesitation. The decelerated words were characterised with the following variables: number of syllables, position of the accent-group in the phrase, presence of an accent, word class. Their role in the information structure was determined in discussion of the two authors, considering the wider context of each utterance.

3.3 Results

The two approaches applied for selecting locally decelerated words yielded considerably different results, since only 19 of 64 words in each list (30%) were identified by both methods. Panel A of Figure 2 shows that the slowest content words were 20–25% slower in relation to the mean local articulation rate of the phrase (with a few outliers decelerated even by 35%). The LAR values of the slowest phones were approximately 10 percentage points lower, because they were considered as individual extremes (unaffected by other phones in the word with faster articulation rate). It is worth mentioning that in the material as a whole, the interquartile range of mean local articulation rate in words before the phrase-final accent-group was from 93 to 105% of the phrase mean.

The majority of words provided by the WORD method were disyllabic, followed by some monosyllabic and a few trisyllabic words. However, the distribution was very different for the PHONE method (see Panel B of Figure 2) – although disyllabic words were also the most common, the other frequent word types consisted of three or four syllables and the list even included five words with 5 to 7 syllables. These results suggest that the WORD method is biased towards shorter words. The choice of method could be also related to the nature of deceleration. If all phones and syllables in a given word were slowed down evenly, both methods should yield the same results. On the other hand, if deceleration was concentrated on a single syllable or even phone, it would be easily detectable with the PHONE method, but its effect on the word mean LAR (assessed by

the WORD method) would diminish with the increasing number of syllables in a word. The present results contained both variants, but the evenly decelerated words were less common (see examples in Figure 3 below). Moreover, Czech content words tend to have 2 to 4 syllables, so the distribution created by the PHONE method seems to reflect these typical numbers more accurately than the WORD method.

Regarding their position in the phrase, target words were frequently found in the initial (31%) or pre-final (45%) accent-group, with only minor differences between the two methods. The remaining 25% of phrases contained the decelerated word in one of the medial accent-groups. If we exclude phrases with three accent-groups (which only have the initial, pre-final and final accent-group), the ratio of the decelerated words in phrase-medial position rises to approximately one third of cases, but the pre-final position remains the most common one.

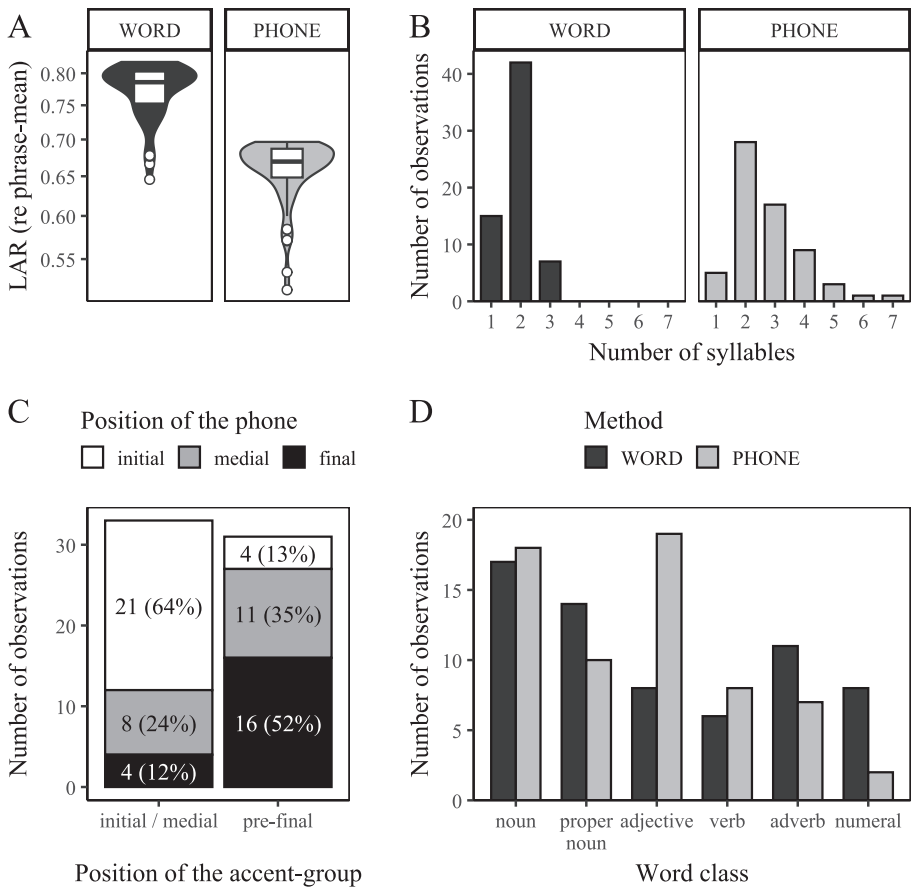


Figure 2 Summary statistics of the selected decelerated content words identified by the WORD and PHONE methods (see text for explanation). Panel A: local articulation rate (LAR) values of the target word or slowest phone. Panel B: target word length (in syllables). Panel C: position of the slowest phone in the target word vs. the accent-group position in the phrase. Panel D: distribution of word classes among the target words.

In Czech, stress falls on the first syllable. In order to avoid stress-clash (accents realised on two neighbouring syllables), Czech monosyllabic words often form an accent-group together with another word, in which only one of them bears the accent. However, stress-clash is sometimes used for emphasis. The accentual status of the monosyllabic target words was checked to see how closely it is related to local deceleration. In both methods, 60% of the monosyllabic words represented cases of stress-clash, while the rest were part of a larger accent-group.

The PHONE method allowed for the analysis of one additional phonetic variable, namely the position of the slowest phone in the decelerated target word. Phones were labelled as either initial (first phone), medial, or final (last phone). Previous research has suggested that prominence introduces decelerations word initially, while final lengthening is known to affect mostly the final one or two syllables in a word (Campbell, 2000, p. 323). In the present data, all phone positions were attested in comparable numbers overall, but a different picture emerged when considering the pre-final accent-groups separately from the others. Panel C of Figure 2 shows that word-final phones were decelerated mostly in pre-final accent-groups, where they amounted to about half of all cases. Instead of marking prominence, these decelerations might be a result of a wider final-lengthening, which extended beyond the final accent-group.

On the other hand, deceleration in initial or medial accent-groups affected mainly word-initial phones. Moreover, 5 of 8 cases of word-medial decelerations also represented phones in the first syllable. The remaining 4 word-final phones could not be explained with final-lengthening, but they could be understood as anticipating prominence on the following syllable. Keeping in mind that the LARometer has a very fine resolution (on the level of phones), the precise position of the temporal prominence 'peak' might be slightly misaligned to the segmental string. Targeting the first syllable of a word could therefore result in the lowest measured LAR values being found anywhere in the interval from the end of the previous syllable to the beginning of the following one.

All classes of content words were represented among the decelerated target words (see Panel D of Figure 2). A special category of proper nouns was introduced due to their reasonably high occurrence. They included names of people, geographical areas, political parties etc., which are often less predictable from the context than other word classes. Comparing the two methods, there were marked differences in the number of adjectives and numerals, which can be related to their number of syllables. Czech is an inflectional language and its adjectives usually consist of multiple syllables (combining morphemes with lexical and grammatical meaning). These were dispreferred by the WORD method, unlike numerals, which are typically mono- or disyllabic.

Although the analysis from the perspective of information structure was also exploratory, there were certain initial assumptions regarding its results. Since the thematic parts of utterances include concepts that are already present in the shared knowledge of the communication partners, it was expected that the decelerated words would more likely have rhematic roles, which convey new and often unpredictable information.

A few examples of rhematic decelerated words are presented in Panels A, B and C of Figure 3. The first two phrases formed one sentence – '*According to him, the document creates a* ***new*** *European state | which reduces the* ***influence*** *of individual member states.*' They followed a sentence stating that a politician rejected the proposal for a European

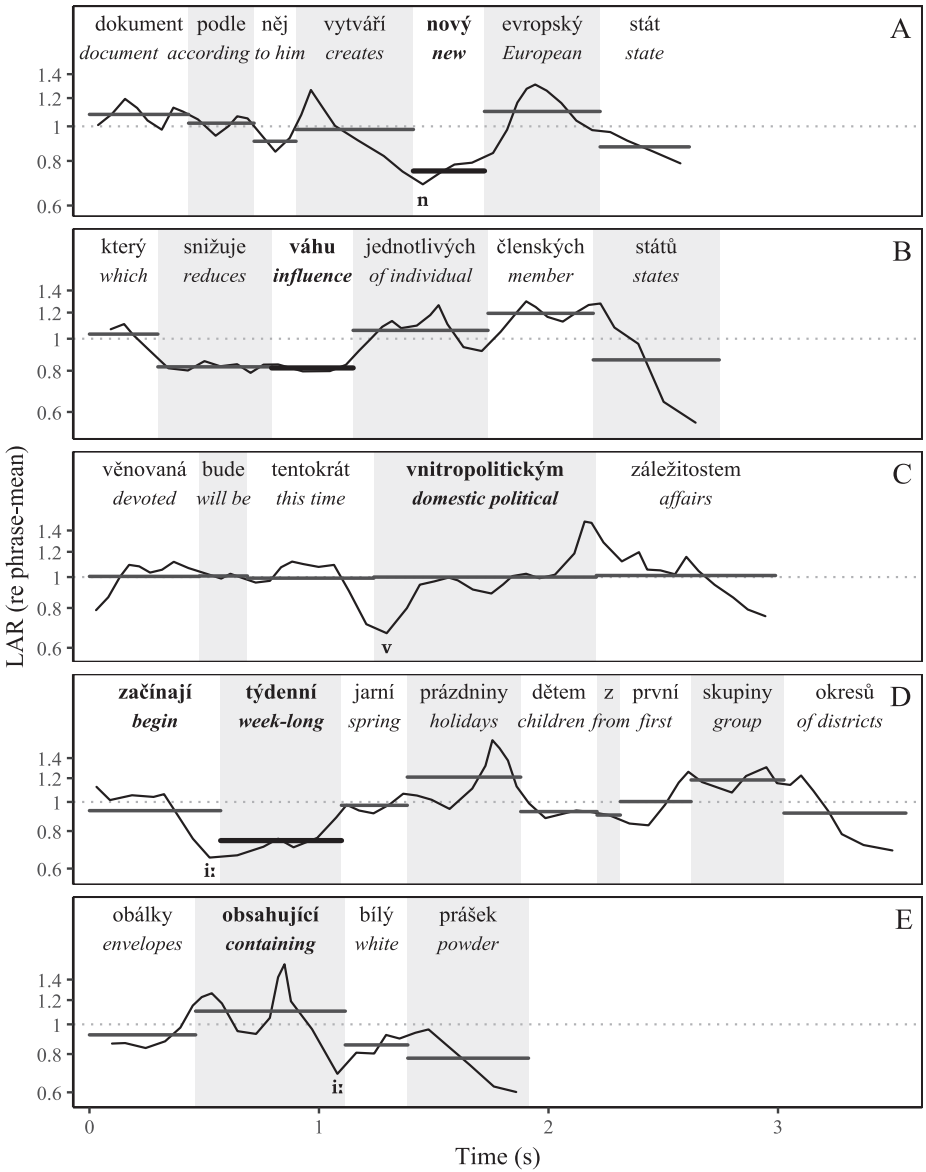


Figure 3 The local articulation rate (LAR) contours (smoothed with a 3-point moving average) of selected phrases containing locally decelerated words. The horizontal lines represent LAR means in words; the decelerated target words are highlighted in bold. For wider contexts and interpretations (especially for the phrase in Panel E), see text.

constitution. The references to the politician ('*him*') and the European constitution ('*the document*') are therefore part of the theme, while the decelerated word '*new*' belongs to the rheme. The crucial concept in the second phrase is the reduction of influence, since member states are implicitly referred to by mentioning the European constitution, which

is linked to the European union. Although the word '*influence*' was identified as the target decelerated word, the LAR contour shows that the other important word '*reduces*' was decelerated to the same extent (by nearly 20% from the mean LAR of the phrase).

While the target word in Panel A was found by both methods, the word in Panel B was only found by the WORD method. The phrase in Panel C contains another rhematic word ('*This time it will be devoted to* ***domestic political*** *affairs.*'), however, this one was considered as decelerated only by the PHONE method. The LAR contour indicates that only the first few phones were slowed down and the final (sixth) syllable was articulated even faster than the middle part of the word. Decelerated words in Panels B and C thus nicely illustrate the shapes of LAR contours (in words) that the two methods aim at.

Table 3 shows that the rhematic decelerated words prevailed. However, the quarter or even third of cases (depending on the method) which belonged to the theme needs to be explained, keeping in mind that these target words were also markedly locally decelerated. A more detailed categorisation of the information structure roles distinguished contrastive topics as a subtype with a more prominent function, but this role applied only to a few words. One such case identified by the WORD method is presented in Panel D of Figure 3, in the sentence '*(…) the* ***week-long*** *spring holidays begin for children from the first group of districts.*' The previous context discussed the types of holidays that await pupils in the near future, and the specification of '*week-long*' contrasted them with '*one-day holidays*' mentioned directly beforehand. Unlike the WORD method, the PHONE method pointed to the first word of the phrase ('*begin*'), because its final vowel was the slowest phone of the phrase. This would be a typical case of the deceleration anticipating the prominence of the following syllable that was discussed earlier. This shift was in fact caused by the smoothing – according to the original LAR values, the initial [t] in the word '*week-long*' was actually slower than the previous phone [i:].

Table 3 Number of decelerated words according to their role in the information structure of the utterance. The potentially more prominent roles are coloured in grey.

	Role in the information structure	Number of cases (%)	Role in the information structure	Number of cases (%)
WORD	theme	18 (28%)	theme proper	2 (3%)
			part of theme	13 (20%)
	rheme	46 (72%)	contrastive topic	3 (5%)
			part of rheme	37 (58%)
			rheme (contrast)	1 (2%)
			rheme proper	8 (12%)
PHONE	theme	23 (36%)	theme proper	---
			part of theme	21 (33%)
	rheme	41 (64%)	contrastive topic	2 (3%)
			part of rheme	34 (53%)
			rheme (contrast)	1 (2%)
			rheme proper	6 (9%)

Out of the 21 thematic target words identified by the PHONE method (excluding the contrastive topics), 10 were selected due to the deceleration of the word-final phone. Some of them were anticipating the prominence of the following syllable like the example in Panel D, while others were part of the pre-final accent-group, which could be a result of a wider phrase-final deceleration. Approximately a third of the thematic words found by the WORD method were proper names. The speakers could possibly have considered these as inherently prominent due to their low predictability from the context. There were also adjectives which were adding new information to the respective nouns, despite belonging to the theme. For example, in the utterance '*Unknown perpetrators send envelopes containing white powder to members of government and parliament.*', the words '*envelopes*' and '*powder*' were repeated from the previous sentence, which mentioned '*envelopes with suspicious powder*'. The specification of the colour therefore extends the topic. Note that the slowest phone was also found in the word preceding the adjective '*white*' (see Panel E in Figure 3).

3.4 Discussion

The present study illustrated that the LARometer is capable of capturing phrase-internal variations in articulation rate, as well as phrase-final deceleration (present in all contours in Figure 3). The analysis of some phonetic variables highlighted the differences between the two approaches in identifying decelerated words. The PHONE method was found to be less biased towards shorter words. This suggests that prominent words do not have to be decelerated throughout – slowing down in just part of them functions as well. It follows that the position of the decelerated phone in the target word needs to be taken into consideration while interpreting the results.

Most decelerated words conformed to the assumption that they would have prominent roles in the information structure of the utterances, but the overall picture was not clear-cut. A partial explanation may lie in the time constraints on the production of radio news. Informal observations suggest that genres which are more spontaneous or less time-constrained time might provide greater variability and larger changes of local articulation rate, which might be more directly linked to the information structure. Furthermore, it was observed that the nature of news presenting leads to texts with very high information density. Instead of working with whole utterances, these texts would deserve a finer analysis of the functional and prominence relations between words on a lower level.

4. Study 2: Differences in the realised and canonical articulation rate in spontaneous speech

4.1 Aims

Section 2.4 introduced two metrics that the LARometer can compute – the realised and canonical local articulation rate. In clear speech, their values are mostly the same, but other speech styles can exhibit reductions and elisions, which would manifest as differences between the two measures. The second study was conducted in order to show the use of the LARometer for identifying highly reduced words.

4.2 Method

A corpus of television political debates was used as the material for this study. It consisted of extracts from 16 speakers (all men) in the length of at least 500 words per speaker. The debates featured one or two politicians and a moderator (who was one of the 16 speakers). The extracts were chosen from the middle part of each debate (excluding the first and last 10 minutes of the debate, which on average lasted one hour). There were 8,871 words in total, corresponding to 59 minutes of speech.

The TV debates are a dialogical genre (even though the roles of the moderator and the guest are asymmetrical). The participants are forced to defend their opinions or even argue with a political opponent, and they usually have to react quickly and without preparation. As a result, this material was expected to contain a reasonable number of reductions, since they are linked to natural continuous speech processes and they frequently occur in more spontaneous and informal styles. The debates also exhibited many moments of more people speaking at the same time, however, these overlapping utterances were not included in the sample.

All recordings were annotated on the level of phones and phonemes, and the placement of phone boundaries was manually corrected. Both the realised and canonical local articulation rate (LAR) values were calculated for all utterances. The values were smoothed with a 3-point moving average, and means of realised and canonical LAR were obtained in words. Subsequently, we calculated the difference between the canonical and realised word mean values for each word. The words with higher canonical than realised LAR were then analysed, focusing on the extent of the rate difference and the types of words that were reduced.

4.3 Results

Overall, the material contained 784 words with reductions detectable by the described method. Figure 4 shows the histogram of the measured differences in word-mean canonical and realised local articulation rate. The canonical LAR values were typically approx-

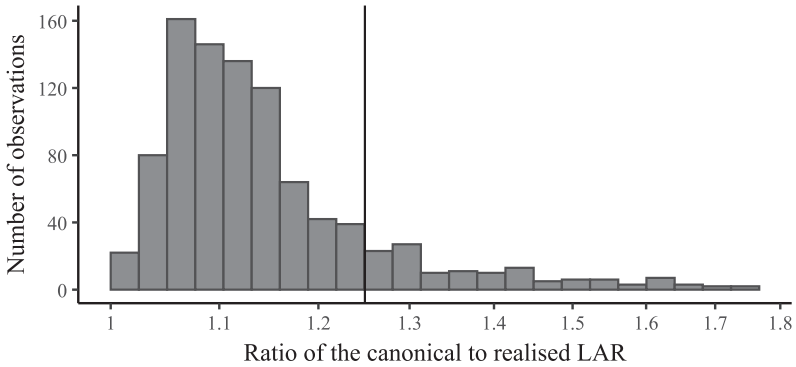


Figure 4 The histogram of the differences between the two local articulation rate (LAR) measures, expressed as the ratio of the canonical to the realised mean LAR in words. For 16% of words, the ratio was higher than 1.25 (indicated by the vertical line).

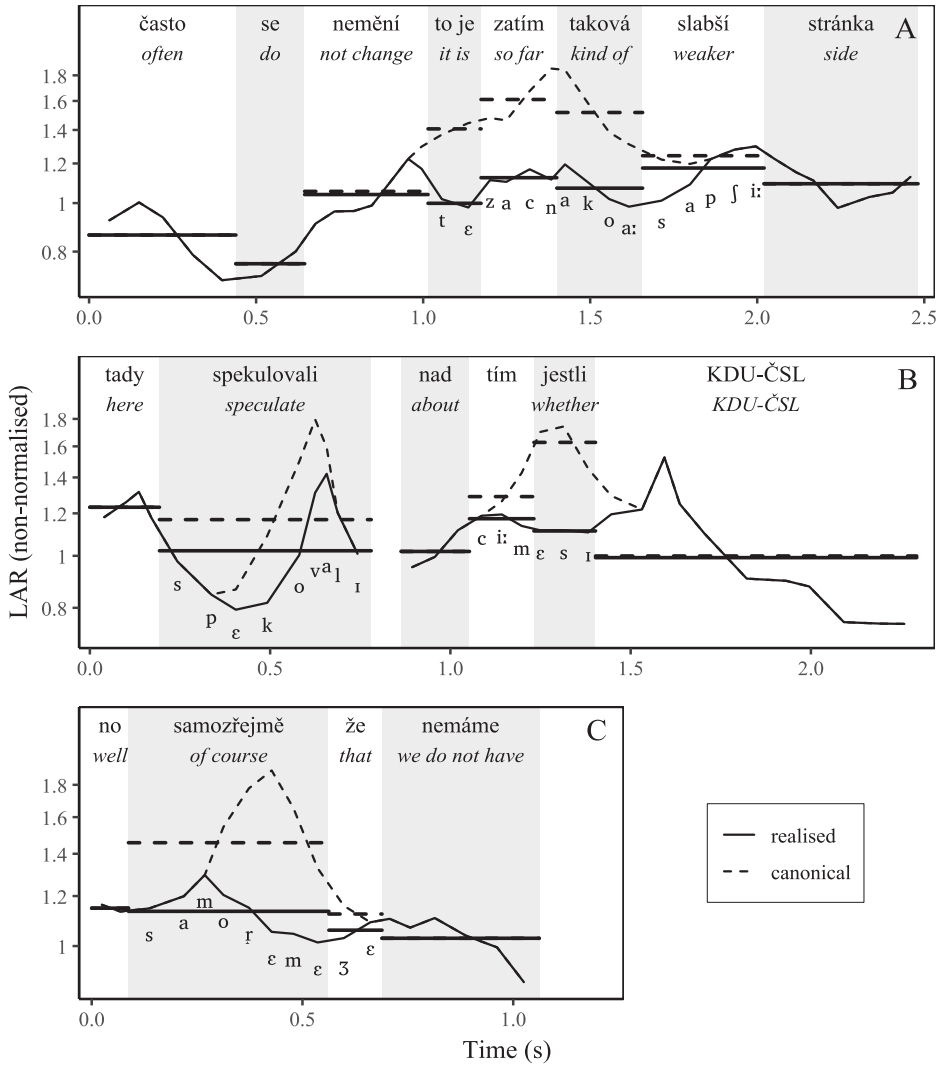


Figure 5 The realised (solid line) and canonical (dashed line) local articulation rate (LAR) contours of selected phone reductions, smoothed with a 3-point moving average. The horizontal lines represent realised and canonical LAR means in words.

imately 5–10% higher as opposed to the realised LAR. In some words (on the left side of the histogram), the differences only amounted to a few percent. These were often caused by phone alternations, e.g. changes in consonant voicing or manner of articulation. Nevertheless, the measure was also affected by the word's length – due to the averaging, the elision of a phone manifests as a smaller difference in longer words.

The distribution was clearly right-skewed and 16% of words had ratios of the two measures between 1.25 and 1.80 (located on the right side of the vertical line in Figure 4). This means that if the canonical form of each word was taken into account, their local

articulation rates would be 25% to 80% higher in comparison with the rates based only on the articulated phones. These 122 significantly reduced words (for simplicity called 'outliers') were thus analysed further and they were contrasted with the whole sample. In general, most of the reduced words were content words (73%) and the rest were function words (27%). However, the relationship was reversed for the outliers, with the function words (67%) outnumbering the content words (33%). The result can be explained as a combination of two factors. Firstly, function words are more common and predictable, which makes them prone to reductions without decreasing intelligibility. Secondly, they are usually shorter, so any reduced or elided phone has a strong effect on the difference between the mean canonical and realised LAR.

In fact, the most frequent case among the outliers was a combination of two words merged into one – '*to je*' ('it is') pronounced as [tɛ] instead of the full form [to jɛ]. Both of these words are semantically very vague and listeners can infer them from the context. They often function as a reduced thematic part or signal a paraphrase. In the Panel A of Figure 5, this elision of two phones from the canonical four led to a 41% higher canonical LAR in relation to the realised LAR. The speaker reduced also the following two words in this utterance (to a similar extent). He elided the second vowel of the word [zaciːm] and merged its final nasal with the initial plosive [t] of the next word, producing [n] instead of [mt]. One elision (of the phone [l]) was also present in the word '*slabší*' [slapʃiː].

Other recurrent function words among the outliers included '*jestli*' ('if, whether'), '*protože*' ('because') and '*je*' ('is' or 'them'). The word '*jestli*' [jɛstlɪ] often undergoes a simplification of the consonant cluster through the elision of [t], however, the following [l] was also elided in most realisations in the analysed material. Moreover, '*jestli*' was sometimes shortened even to [ɛsɪ], as in the example in Panel B. Since the contours show smoothed values, the difference between the realised and canonical LAR spreads also to the neighbouring words.

The list of reduced content words was more variable, but two words were subject to strong reduction more often than others – '*samozřejmě*' ('of course, obviously') and '*šest*' ('six'). The first one is often used as a marker modifying the main message and phonetic reductions accompany this transition from a content word with a specific meaning towards a function particle. The numeral 'six' (pronounced [ʃɛst]) contains a consonant cluster, which is frequently simplified (especially in a less formal style), since listeners are not likely to confuse it with other numerals. Panel C in Figure 5 presents an example of the word '*samozřejmě*' [samozr̝ɛjmɲɛ] reduced to [samor̝ɛmɛ]. Panel B contains another reduced content word – the phones [ul] in the middle of the verb '*spekulovali*' [spɛkulovalɪ] were elided. Although this resulted in a loss of one syllable, the mean word canonical LAR was only 14% higher than the realised LAR, probably due to the larger number of phones in the word. This value thus represents a more typical difference between the two LAR measures in the material.

All reduced words were also analysed with respect to the phones that were substituted or elided. The most frequently reduced phone was clearly the approximant [j], which alone accounted for 19% of all instances of reduced phones. Together with [l t ɦ o d v] (in descending order of frequency), these seven phones represented 63% of all phone reductions. The consonants [j l ɦ] were typically elided intervocalically, while [t d v] more likely in consonant clusters. The presence of the vowel [o] among the most reduced phones was

due to the words *'to je'* (as described earlier) and '*protože*' ([protoʒɛ], often pronounced as [pr̩toʒɛ] or [pʒɛ]).

Some individual differences could be found between the 16 speakers regarding the number of significantly reduced words. The median count was 7 such words in the approximately 500-word extract. There was one speaker with a clear speaking style (only 1 significantly reduced word). On the other hand, two speakers produced 16 words with the word-mean canonical LAR 25–80% higher than the realised LAR. Interestingly, these differences could not be explained by the average local articulation rate of these speakers, since the clear speaker was the second fastest and the two reducing speakers had medium articulation rates.

4.4 Discussion

The present study showed that it is possible to identify reduced words with the two measures computed by the LARometer. Moreover, the degree of reduction can be quantified. The current results were based on mean values in words, since these represent meaningful units of speech. However, one could adopt a different approach and compare the canonical and realised LAR per phone or syllable – the extent of the ratios would not be affected by the length of the word they are a part of.

Apart from describing the words and phones that were most frequently subject to reduction, the results also showed some speaker-individual differences. Quite remarkably, the number of reductions was not dependent on the mean articulation rate. The material of these political debates could therefore serve as a source of speech samples with various combinations of fast/slow and clear/casual speech, which could be used in experiments on the perception of speech tempo and formality.

5. Conclusion

The article has introduced the LARometer tool for calculating the local articulation rate (LAR) in speech. Unlike simple rate measures, which are expressed in phones or syllables per second, the LAR is a dimensionless unit. It relates the observed durations of speech segments to their inherent durations. This normalisation reduces the effects of individual phones' typical duration and allows for a very local perspective on the changes in articulation rate. At the same time, mean LAR values can be compared across words or phrases containing different numbers and types of phones. The LARometer should therefore capture prosodically relevant changes in articulation rate with potential communicative meaning. However, its perceptual validity needs to be tested as the next research step in order to see whether the calculated changes in local articulation rate correspond to differences perceived by listeners. Experiments could also try to determine the size of perceivable local articulation rate changes, since previous research on just noticeable differences was mostly concerned with global rate differences (e.g., Quené, 2007).

We also presented two studies, which illustrated the application of the proposed approach on authentic speech material. They focused on the relationship of the local

articulation rate with information structure of utterances, and on phone reductions, although many more research questions could be asked and explored with the metrics provided by the LARometer. While the results seemed promising and meaningful, they were only preliminary. Future research could test some of the suggested hypotheses with more data and with statistical analyses.

Furthermore, temporal aspects of speech should not be considered independently of other prosodic domains. The quantification of local articulation rate enables it to be analysed jointly with F0 contours or intensity contours (cf. Campbell, 1992). The LARometer could also be applied to other languages than Czech. Although the inherent durations of phones might be language specific, the principles of normalisation used in the LARometer work universally, provided there is a sufficient labelled speech corpus available for the given language.

Acknowledgements

This work was financially supported by the Charles University Grant Agency, project no. 32424, entitled 'Deceleration as a marker of prominence in the speech of newsreaders and in political debates', implemented at the Faculty of Arts, Charles University.

REFERENCES

Baumann, S., Becker, J., Grice, M., & Mücke, D. (2007). Tonal and articulatory marking of focus in German. *Proceedings of the XVIth ICPhS*, 1029–1032.

Beckman, M. E., & Ayers Elam, G. (1993). *Guidelines for ToBI labelling*. The Ohio State University Research Foundation.

Boersma, P., & Weenink, D. (2024). *Praat: doing phonetics by computer* (Version 6.4.04) [Computer software].

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.

Campbell, N. (1992). Prosodic encoding of English speech. *2nd International Conference on Spoken Language Processing (ICSLP 1992)*, 663–666.

Campbell, N. (2000). Timing in speech: a multi-level process. In M. Horne (ed.), *Prosody: theory and experiment* (vol. 14, pp. 281–334). Springer Netherlands.

Cooper, W. E., Eady, S. J., & Mueller, P. R. (1985). Acoustical aspects of contrastive stress in question–answer contexts. *The Journal of the Acoustical Society of America*, *77*(6), 2142–2156.

Heldner, M., & Strangert, E. (2001). Temporal effects of focus in Swedish. *Journal of Phonetics*, *29*(3), 329–361.

Klatt, D. H. (1975). Vowel lengthening is syntactically determined in a connected discourse. *Journal of Phonetics*, *3*(3), 129–140.

Koreman, J. (2006). Perceived speech rate: the effects of articulation rate and speaking style in spontaneous speech. *The Journal of the Acoustical Society of America*, *119*(1), 582–596.

Local, J. (1992). Continuing and restarting. In P. Auer & A. Di Luzio (eds.), *Pragmatics & Beyond New Series* (vol. 22, pp. 273–296). John Benjamins Publishing Company.

Machač, P., & Skarnitzl, R. (2009). *Fonetická segmentace hlásek* (1st ed.). Epocha.

Miller, J. L., Grosjean, F., & Lomanto, C. (1984). Articulation rate and its variability in spontaneous speech: a reanalysis and some implications. *Phonetica*, *41*(4), 215–225.

Pedersen, T. L. (2024). *patchwork: the composer of plots* [Computer software].

Plug, L., Lennon, R., & Smith, R. (2022). Measured and perceived speech tempo: comparing canonical and surface articulation rates. *Journal of Phonetics*, *95*, 101193.
Pollák, P., Volín, J., & Skarnitzl, R. (2007). HMM-based phonetic segmentation in Praat environment. *The XII International Conference Speech and Computer – SPECOM 2007*, 537–541.
Quené, H. (2007). On the just noticeable difference for tempo in speech. *Journal of Phonetics*, *35*(3), 353–362.
R Core Team. (2024). *R: a language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing.
Saarni, T., Hakokari, J., Isoaho, J., & Salakoski, T. (2008). Utterance-level normalization for relative articulation rate analysis. *Interspeech 2008*, 538–541.
Šturm, P., & Bičan, A. (2021). *Slabika a její hranice v češtině*. Karolinum.
Trouvain, J. (2003). *Tempo variation in speech production*. [Doctoral dissertation, Saarbrücken University]
Uhmann, S. (1992). Contextualizing relevance: on some forms and functions of speech rate changes in everyday conversation. In P. Auer & A. Di Luzio (eds.), *Pragmatics & Beyond New Series* (vol. 22, pp. 297–336). John Benjamins Publishing Company.
Wagner, P., Trouvain, J., & Zimmerer, F. (2015). In defense of stylistic diversity in speech research. *Journal of Phonetics*, *48*, 1–12.
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.

RESUMÉ

Článek představuje metodu normalizace inherentních temporálních vlastností hlásek (LARometr), která poskytuje relativní míru lokální artikulační rychlosti (LAR). Tato metoda umožňuje kvantifikaci komunikačně relevantních změn artikulační rychlosti a jejich zobrazení v podobě rychlostních kontur. Normalizace je založena na rozsáhlém ručně anotovaném korpusu obsahujícím více než čtyři hodiny souvislé řeči. Použití LARometru je ilustrováno na příkladu dvou studií. V rámci první studie byla v českých rozhlasových zpravodajstvích vyhledána lokálně zpomalená plnovýznamová slova. Tato zpomalená slova měla často prominentní role z hlediska aktuálního členění (réma, kontrastivní téma). Výsledky také ukázaly, že zpomalování zasahuje různé části slov. Druhá studie se zaměřila na hláskové redukce v televizních politických debatách. Redukce se nacházely především v plnovýznamových slovech, nicméně neplnovýznamová (gramatická) slova byla redukcemi ovlivněna výrazněji. Jednotliví mluvčí se také značně lišili množstvím produkovaných redukcí.

Michaela Svatošová
Institute of Phonetics
Charles University, Faculty of Arts
Prague, Czech Republic
michaela.svatosova@ff.cuni.cz

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

2025 ACTA UNIVERSITATIS CAROLINAE PAG. 33–42
PHILOLOGICA 3

PERCEPTUAL EVALUATION OF THE EFFECT OF EXTERNAL RADIOTHERAPY IN THE NECK AREA ON CHANGES OF VOICE AND THE VOICE QUALITY OF CZECH PATIENTS

JAN ŠEBEK, TOMÁŠ BOŘIL

ABSTRACT
This research is focused on changes of voice and the voice quality before radiotherapy treatment and in time of one year after external radiotherapy (RT) in the neck area of Czech patients and to compare measurement for different subgroups of patients (by age, gender, surgical resection, aphonia, etc.). The perceptual test was performed on 16 patients undergoing external radiotherapy in the neck area and the changes of voice and the voice quality compared at before RT and 1, 6 and 12 months after RT. 2 clinicians and 1 trained voice specialist evaluated GIRBAS parameters of the voice quality and 96 lay listeners evaluated the scope of change of voice. The results of perceptual tests of lay listeners point to the difference of change in voice is the most pronounced in patients who had aphonia during the RT. Of the GIRBAS parameters, instability and roughness changed the most after RT treatment.

Keywords: changes of voice; voice quality; perceptual test; GIRBAS parameters; radiotherapy treatment; Czech patients

1. Introduction

In the last 20 years, there has been an increasing number of patients in the Czech Republic who had to undergo RT in the neck area as a part of their curative treatment of oncological diseases (Krejčí et al., 2022). One of the side effects of this treatment may be a temporary or permanent change of voice and the quality of the patients' voice and speech (Dršata et al., 2008).

Several research studies on this issue have been carried out internationally so far (Lazarus, 2009), but all of them were conducted especially from a medical perspective, focusing only on basic phonetic, especially acoustic, parameters (fundamental frequency, standard deviation of the fundamental frequency, jitter, shimmer, harmonicity, etc.) (Bagherzadeh et al., 2022; Bibby et al., 2007; Fung & Yoo, 2001), or questionnaire methods were used to examine the impact on the quality of life of patients after radiotherapy treatment (Mekiš et al., 2022; Šiupšinskienė et al., 2008) or some foreign studies focused on differences in voice quality when completing or not completing voice rehabilitation after radiotherapy treatment (Mei-jia et al., 2018; Tuomi et al., 2014).

https://doi.org/10.14712/24646830.2025.19

In addition to studying changes in acoustic parameters, it was also appropriate to focus research from a perceptual perspective. The voice quality assessment was performed using the GIRBAS method, which had become the most widely used voice scaling method by clinicians (Uloza et al., 2005; Karnell et al., 2007). The GIRBAS scale for subjective voice evaluation containing six voice-quality parameters, G (Grade of dysphonia), I (Instability), R (Roughness), B (Breathiness), A (Asthenia), and S (Strain), meet the criteria of reliability and relevance well, and the scale is easy to use (Olivares et al., 2023). Yamauchi et al. (2010) state that S parameter can be defined as extent to which strain or hyperfunctional use of phonation is heard, A parameter as degree of weakness heard in the voice, B parameter as degree to which air escaping from between the vocal folds can be heard by examiner, R parameter as impression of irregularity of the vibration of the vocal folds and I parameter as degree of change of the voice quality over time.

In the Czech environment, changes in the voice of patients before and after RT have not been yet evaluated by perceptual tests completed by lay listeners who, during the listening test, did not know that they were hearing the voices of patients before and after RT. From this perspective, it is beneficial to find out how the changes in the voice of these patients are perceived by public, whether the changes of voice are audible, and whether the lay listeners perceive the patient's voice before treatment compared to the voice after treatment as the voice of the same speaker.

The main aim of the present study was to investigate the effect of RT treatment on voice and evaluated the changes of voice and the voice quality perceptually. The changes of voice and the voice quality compared at before RT and 1, 6 and 12 months after RT. On these results analyse if they indicate any trend or are correlated to any properties of research group.

2. Method

2.1 Research group

The study was started after obtaining approval from the ethics committee of Bulovka University Hospital (No. 12.10.2021/10214/EK-Z). 16 respondents (4 women and 12 men) aged 19 to 76 years (average age 56 years) voluntarily participated in the research study and successfully completed the entire research study. All respondents were native Czech speakers, as required by the research study. All respondents at the Department of Radiation Oncology of Bulovka University Hospital in Prague underwent external radiotherapy treatment of the neck area as a part of their curative treatment of carcinoma in larynx area. The research group consisted of 2 smokers and 14 non-smokers, 5 of whom reported that they had been smokers in the distant past (more than 5 years ago). All respondents received a total radiation dose to the tumor site ranging from 66 to 77 Gray (Gy). 11 respondents also received curative radiation to the site of metastasis (neck node area), with a total dose of 54 to 66 Gy. 15 respondents underwent radiotherapy treatment in 33 fractions and 1 respondent in 35 fractions. All respondents received external radiotherapy to the neck area using the VMAT (Volumetric Modulated Arc Therapy) method

with IGRT (Image-Guided Radiation Therapy) on a linear accelerator with coverage of areas outside the planning target volume using a multi-leaf collimator in the normofractionation mode (5 fractions/week with a dose of up to 10 Gy/week). 4 respondents underwent surgical resection of part of the tongue before participating in this study, and 6 respondents had up to 2 weeks of aphonia during RT treatment.

2.2 Control group

The control group consisted of 4 Czech native speakers (2 women and 2 men) aged 38 to 52 years (average age 46 years) who were recruited at the start of this study. The control group was matched for age and gender. These speakers were representatives of a healthy population and had not any vocal abnormalities or diseases. Voice recordings of members of the control group were used to test the ability of lay and educated listeners to objectively evaluate the voice change and the voice quality of healthy and diseased voices of speakers.

2.3 Material

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 48 kHz sampling frequency and 16-bit resolution. Four recording sessions were analysed in this study – first recording before RT, second in time of 1 month after RT, third at 6 months after RT and the last (fourth) at 12 months after RT. 20 tasks were recorded in each session with each respondent and once with each member of control group. Three of all tasks were used for perceptual test: spontaneous speech about some experiences of the last days (the first task, approx. 1 minute long), reading text of 5 sentences (the fourteenth task, see appendix) and sustained vowel /a/ (the nineth task).

2.4 Test of GIRBAS parameters

In order to evaluate the vocal competence of the patients before and after the RT treatment, the GIRBAS perceptual assessment scale were used to evaluate their voice quality. The evaluators evaluated voice quality of all respondents and all members of control group. Same set of three recordings (spontaneous speech about some experiences of the last days, reading text of 5 sentences and sustained vowel /a/) was used to evaluate each GIRBAS parameter.

The evaluation was made by 2 clinicians from Department of Phoniatrics at General Faculty Hospital in Prague and 1 trained voice specialist from Institute of Phonetics in Prague. The evaluation was made by giving a score from 0 to 3 (0 = normal, 1 = mildly affected, 2 = moderately affected, 3 = highly affected) including half grades for a finer scale of evaluation (0,5 = value between normal and mildly affected, 1,5 = value between mildly affected and moderately affected, 2,5 = value between moderately affected and highly affected).

Evaluators were allowed to replay any of the recordings. To minimize the order effect and also achieve the most objective assessment of the voice quality possible, evaluators did not know if the set of three recordings was recorded by member of research group or

member of control group and also did not know in which time of treatment the recordings were recorded. Sets were randomized and evaluators evaluated 4 sets weekly in half of year. Their ratings correlated with each other, so all data were used for analysis.

2.5 Listening test for lay listeners

For our perceptual test for lay listeners, we used recordings of reading same text of all respondents and all members of control group. In total, 96 recordings (48 pairs of recordings in 3 line-ups) of up to 20 seconds duration of each recording were used for the test. First 16 pairs of recordings (first line-up) were composed of recording of patient's voice before RT and recording of reading the same text by the same patient in time of 1 month after RT. Next 16 pairs of recordings (second line-up) were composed of recording of patient's voice before RT and recording of reading same text by same patient in time of 6 months after RT. And the last 16 pairs of recordings (third line-up) were composed of recording of patient's voice before RT and recording of reading same text by same patient in the time of 12 months after RT. The volume of the audio recording was unified to the same level for all recordings. The test was designed in PsyToolkit (Stoet, 2017) and it could be filled out at any time between March 15 and April 30, 2025. For each pair of recording, the listener had to decide, using a cursor on a scale, how much the speaker's voice had changed due to the disease, without the listener knowing what the disease was. On the scale, the minimum (value 0) was indicated as "no change in the voice of speaker" and the maximum (value 100) was indicated as "such a change in the voice that it is the voice of another speaker".

If a patient's recording was missing from the pair of recordings (there were 4 cases due to health reasons he was unable to attend the recording of the tasks), this recording was replaced with a recording of the same reading text spoken by the voice of a member of the control group of the same gender to test the ability of listeners to objectively evaluate the difference between healthy and diseased voices of speakers (we expected a rating close to 100 for these pairs of recordings).

Listeners were allowed to replay any of the recordings. In order to minimize the order effect, samples were randomized (in each line-up, as well as the line-ups themselves).

Besides the test, we gathered basic information from the listeners (gender, age, occupation). The test was successfully completed by 68 female and 30 male listeners, aged 20 to 79 (average age 45 years), all coming from the Czech Republic and from varied occupation (unemployed, hairdresser, social worker, courier, teacher, clerk, administrative worker, librarian, waiter, IT specialist, gardener, doctor, student, musician, physiotherapist, businessman, quality manager, HR specialist, designer, pensioner etc.). The test was intended and focused on lay listeners, precisely to get as close as possible to the perception of change of the voice from the perspective of the general public.

It was revealed later that two listeners had followed the test instructions in reverse (their responses reached negative values of the correlation coefficient against the other listeners), and their responses were eliminated. Therefore, 96 listeners' answers were analysed. The average time of the listeners finished the test (including the information questions) was 32 minutes.

2.6 Statistical analysis

Linear model and correlation analysis were conducted using stats package in R (R Core Team, 2024), data processing and visualisation was performed using tidyverse package (Wickham et al., 2019) and scatterpolar plots were plotted using plotly package (Sievert, 2020).

3. Results and discussion

External radiotherapy in the neck area has effect on changes of the parameters G (Grade of dysphonia), I (Instability) and R (Roughness) the most, while RT treatment has minimal effect on B (Breathiness), A (Asthenia) and S (Strain). Compared to the values of the GIRBAS parameters before RT treatment (marked as N1) and with the time interval after the end of RT (1 month after RT marked as N2, 6 months after RT marked as N3 and 12 months after RT marked as N4), the values of the I parameter increase (from value 0.66 to 1.17). The highest value of the G parameter was found in voices recorded with an interval of 6 months after RT treatment (value 1.31) and the highest value of the R parameter was found in voices recorded with an interval of 1 month after RT treatment (value 0.78) (see Table 1). Changes and dispersions of parameter values are depicted in the Figure 1.

Table 1 Average values of GIRBAS parameters.

Groups of respondents / time of recording	G	I	R	B	A	S
All members of research group / before RT (N1)	0.84	0.66	0.58	0.04	0.02	0.18
All members of research group / 1 month after RT (N2)	1.02	0.73	0.78	0.10	0.06	0.20
All members of research group / 6 months after RT (N3)	1.31	1.05	0.71	0.01	0.02	0.23
All members of research group / 12 months after RT (N4)	1.24	1.17	0.51	0.01	0.04	0.14
All members of control group / N/A	0.29	0.17	0.04	0.00	0.00	0.04

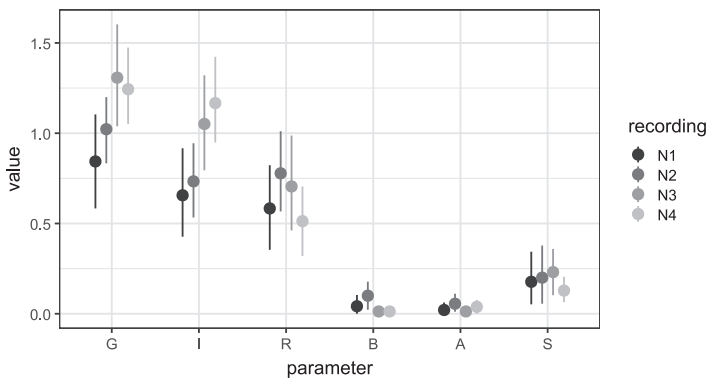


Figure 1 Changes of values of GIRBAS parameters of research group before (N1) and after RT (N2 = 1 month, N3 = 6 months, N4 = 12 months), point = mean value, error bars = 95% confidence interval estimated via a bootstrap method.

Comparing data by age, gender, aphonia, occurrence of metastases and surgical resection, no associations and correlation with the value and its change were found.

If we were to display the IRBAS parameters data using radar graphs, we can compare the graphs to see whether the overall voice impairment is worsening (distance from the origin is increasing) or improving (distance from the origin is decreasing). At the same time, we can determine from the shift of the pentagon shape defined by the IRBAS parameter values whether this shift is, for example, typical for the disease in question or is related to the treatment and its success (see Figure 2).

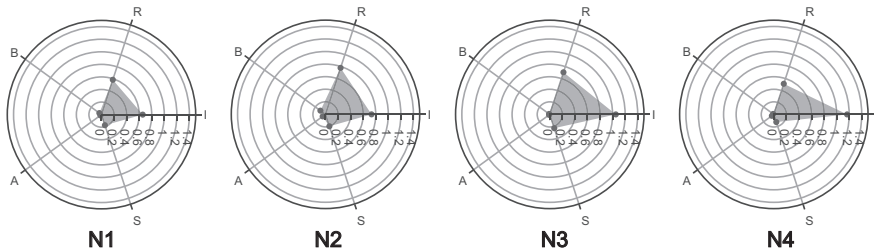


Figure 2 Radar plots of IRBAS pentagons of research group (scale zoomed).

Using different representations of changed values led us to the idea of a model that would most accurately determine the value of the G parameter from the values of other IRBAS parameters. If we assume that all other IRBAS components contribute to the total voice damage (parameter G), then we could determine that the relationship for the value of the total voice damage can be given by the linear model formula

$$G = \alpha + a.I + b.R + c.B + d.A + e.S + \varepsilon$$

where *a, b, c, d, e* are linear coefficients, α is the intercept value and ε is the error therm, the parameter G thus becomes a linear combination of the other parameters. From the actual IRBAS parameter values of this study, we obtained the formula as the most accurate relation for the G parameter:

$$G = 0.629I + 0.467R - 0.056B + 0.385A + 0.030S + \varepsilon$$

The found formula shows that the I, R, and A parameters contribute the most to the G parameter and the B and S parameters the least. Pearson correlation between actual values and the model predictions is 0.896.

Another approach to estimation of the G parameter is an Euclidean distance,

$$G = \sqrt{I^2 + R^2 + B^2 + A^2 + S^2} + \varepsilon$$

where ε is the error therm. Pearson correlation between actual values and the model predictions is 0.940.

The perception of lay listeners also confirmed that voice changes occur as a result of RT treatment. Based on data from a listening test for lay listeners, it was found that apho-

nia determines how much a change in voice will be perceived by the listener. The voices of patients who did not experience treatment-related aphonia were rated as having less change, while the voices of patients who experienced treatment-related aphonia were perceived as having more change and, in extreme cases, were rated as voices of different speakers (see Figure 3).

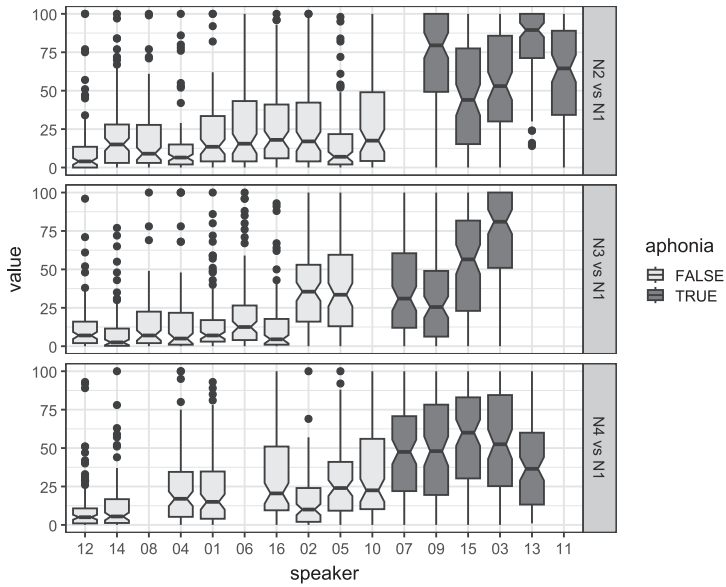


Figure 3 Evaluation of voice changes by patients according to aphonia.

Comparing results of listening test for lay listeners by age, gender, occurrence of metastases and surgical resection, no associations and correlation with the value and its change were found.

During the statistical processing of the data, we found that for the question of how much the listener perceives the change of the voice in a pair of recordings, it was not necessary to use such a fine scale of values as we used (0 to 100) because for the voices of patients

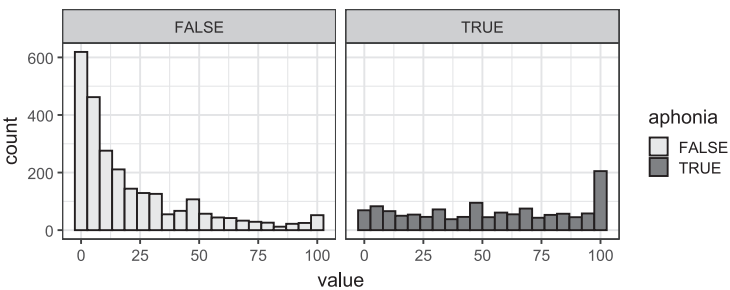


Figure 4 Histogram of listeners' values of voice changes (speakers without and with aphonia). The peaks around 0, 50 and 100 raise the question of whether listeners tend to use the scale in a rather discrete way.

in which there were small changes, listeners tended to give values close to zero, while for the voices of patients in which there were large changes, listeners also tended to give extreme values (close to 100), see Figure 4. Clearly, listeners more often preferred the exact midpoint of the scale (value 50) than the nearby values in both cases (with and without aphonia). It is likely that a discrete scale might be sufficient and more convenient to raters.

4. General discussion and conclusion

In our study, we examined the effect of external radiotherapy in the neck area on changes of voice and the voice quality of Czech patients by perceptual evaluation. It has been perceptually verified that educated and lay listeners perceive changes in the voice and voice quality of patients who have undergone RT treatment. The RT treatment thus most affects voice instability and temporarily roughness. The occurrence of aphonia will affect the perception of the voice change in these patients. In those who had aphonia during the RT treatment, the public perceives the voice change as very significant, audible, and identified as a completely different voice.

In conclusion, this study has shown that when comparing the results of voice recordings of 16 patients, it is not always possible to draw unambiguous conclusions for different groups according to age, gender, etc. However, despite the possible inhomogeneity of the research group, the occurrence of aphonia in these patients was shown to be the main factor of different perceptual evaluation.

Caution should be exercised when formulating conclusions from the analyses: different analysis methods may provide somewhat different results, and the interpretation may therefore be less unambiguous. A higher number of patients involved in the research study would contribute to greater statistical accuracy and more precise modelling.

Acknowledgements

This study was supported by Grant Agency of Charles University, project GAUK no. 26224.

REFERENCES

Bagherzadeh, S., Shahbazi-Gahrouei, D., Torabinezhad, F., Rabi Mahdavi, S., & Salmanian, S. (2022). The effects of (chemo) radiation therapy on the voice and quality of life in patients with non-laryngeal head and neck cancers: a subjective and objective assessment. *International Journal of Radiation Research*, *20*(2), 397–402.

Bibby, J. R., Cotton, S. M., Perry, A., & Corry, J. F. (2007). Voice outcomes after radiotherapy treatment for early glottic cancer: Assessment using multidimensional tools. *Head & Neck*, *30*(5), 600–610.

Dršata, J., Vokurka, J., Čelakovský, P., Hudíková, M., Růžička, J., & Kordač, P. (2008). Přehled foniatrických možností úpravy hlasu po onkologické léčbě nádorů oblasti hlavy a krku. *Onkologie*, *2*(2), 91–93.

Fung, K., & Yoo, J. (2001). Effects of head and neck radiation therapy on vocal function. *The Journal of otolaryngology*, *30*(3), 133–139.

Karnell, M. P., Melton, S. D., Childes, J. M., Coleman, T. C., Dailey, S. A., & Hoffman, H. T. (2007). Reliability of clinician-based (GRBAS and CAPE-V) and patient-based (V-RQOL and IPVI) documentation of voice disorders. *Journal of Voice*, *21*(5), 576–590.

Krejčí, D., Mužík, J., & Dušek, L. (2022). *Novotvary 2019–2021 ČR – Cancer incidence 2019–2021 in the Czech Republic*. Ústav zdravotnických informací a statistiky ČR.

Lazarus, C. L. (2009). Effects of chemoradiotherapy on voice and swallowing. *Current Opinion in Otolaryngology & Head and Neck Surgery*, *17*(3), 172–178.

Mei-jia, Z., Ji-wei, M., Xiang-ru, C., Xin, Z., & Chong, F. (2018). Effect of voice rehabilitation training on the patients with laryngeal cancer after radiotherapy. *Medicine*, *29*, 97–100.

Mekiš, J., Strojan, P., Mekiš, D., & Hočevar Boltežar, I. (2022). Change in Voice Quality after Radiotherapy for Early Glottic Cancer. *Cancers 2022*, *14*(12), 2993.

Olivares, A., Comini, L., Di Pietro, D. A., Vezzadini, G., Luisa, A., Boccali, E., Boccola, S., & Vitacca M. (2023). Perceptual and qualitative voice alterations detected by GIRBAS in patients with Parkinson's disease: Is there a relation with lung function and oxygenation. *Aging Clinical and Experimental Research*, *35*, 633–638.

R Core Team (2024). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org.

Sievert, C. (2020). *Interactive Web-Based Data Visualization with R, plotly, and shiny.* Chapman and Hall/CRC Florida.

Stoet, G. (2017). PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology*, *44*(1), 24–31.

Šiupšinskienė, N., Vaitkus, S., Grėbliauskaitė, M., Engelmanaitė, L., & Šumskienė, J. (2008). Quality of life and voice in patients treated for early laryngeal cancer. *Medicina*, *44*(4), 288–295.

Tuomi, L., Andréll, P., & Finizia, C. (2014). Effects of voice rehabilitation after radiation therapy for laryngeal cancer: A randomized controlled study. *International Journal of Radiation Oncology, Biology & Physics*, *89*(5), 964–972.

Uloza, V., Saferis, V., & Uloziene, I. (2005). Perceptual and acoustic assessment of voice pathology and the efficacy of endolaryngeal phonomicrosurgery. *Journal of Voice*, *19*(1), 138–145.

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., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., & Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686.

Yamauchi, E. J., Imaizumi, S., Maruyama, H., & Haji, T. (2010). Perceptual evaluation of pathological voice quality: a comparative analysis between the RASATI and GRBASI scales. *Logopedics Phoniatrics Vocology*, *35*(3), 121–8.

APPENDIX

The assignment of the fourteenth task (reading text of 5 sentences), which were used for perceptual tests.

Prosím, přečtěte nahlas a zřetelně tento text:

Vítejte na naší frekvenci rádia F.F.U.K. Je tři čtvrtě na sedm a v půl deváté Vás budou čekat zprávy v českém jazyce. Po nich přiklopýtá kolegyně se zajímavostmi ze světa klokanů a kasuárů. Něco nám tu píská, šumí a ševelí. To budou určitě svišti na smetišti.

English translation:

Please read this text loud and clearly:

Welcome to our frequency of radio F.F.U.K. It is a quarter to seven and at half past eight, there will be news in Czech. After that, a colleague will shuffle in with interesting facts from the world of kangaroos and cassowaries. Something is whistling, rustling, and whispering here. It must be the groundhogs in the dumps.

RESUMÉ

Tento výzkum je zaměřen na změny hlasu a kvality hlasu před radioterapeutickou léčbou a v čase jednoho roku po zevní radioterapii (RT) v oblasti krku u českých pacientů a na porovnání změny hlasu a kvality hlasu u různých podskupin pacientů (podle věku, pohlaví, chirurgické resekce, afonie atd.). Percepční test byl proveden s nahrávkami hlasu 16 pacientů podstupujících RT a byly porovnány změny hlasu a kvalita hlasu před RT a 1, 6 a 12 měsíců po RT. Parametry kvality hlasu GIRBAS hodnotili 2 kliničtí lékaři a 1 vyškolený hlasový specialista a rozsah změn hlasu hodnotilo 96 laických posluchačů. Výsledky percepčního testu hodnoceného laickými posluchači ukazují na to, že změna hlasu je nejvýraznější u pacientů, u kterých se během RT vyskytla afonie. Z parametrů GIRBAS se vlivem RT léčby nejvíce změnila nestabilita a drsnost.

Jan Šebek
Institute of Phonetics
Faculty of Arts, Charles University
Prague, Czech Republic
jan.sebek@ff.cuni.cz

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

2025 ACTA UNIVERSITATIS CAROLINAE PAG. 43–60
PHILOLOGICA 3

CUSTOMISING CZECH PHONETIC ALIGNMENT USING HuBERT AND MANUAL SEGMENTATION

ADLÉTA HANŽLOVÁ, VÁCLAV HANŽL

ABSTRACT
This paper presents Prak, a forced alignment tool developed for Czech, with a focus on transparent modular design and phonetic accuracy. In addition to a rule-based pronunciation module and exception handling, Prak introduces a novel application of non-deterministic, backward-processing FSTs to model complex regressive assimilation processes in Czech consonant clusters. We further describe the integration of a HuBERT-based transformer model and training including extensive manually time-aligned data to enhance phone classification accuracy while maintaining ease of installation and use. Evaluation against a manually aligned test corpus demonstrates that the enhanced model significantly outperforms both our earlier Prak-CV model and the long-established previous forced alignment baseline. The new model reduces major boundary errors and mismatches, bringing alignment accuracy closer to manual phonetic segmentation standards for Czech. We emphasize both methodological transparency and practical usability, aiming to support phoneticians working with Czech as well as developers interested in extending the tool for other languages.

Keywords: forced alignment; phonetic segmentation; Czech; HuBERT; Prak; Praat

1. Introduction

The majority of phonetic analyses require labelling recordings to identify their contents (the type of content varying based on the research question) in order to accurately measure properties that are to be related to said contents. Among the most common types of audio labelling is identifying phone boundaries as a means to then measure formant frequencies, spectral or temporal properties, assess pronunciation and much more. Labelling recordings is often a very tedious and time-consuming process, but also one that is vital to ensure the validity of measurements that are made based on the determined time boundaries. The general effort therefore is to automate these processes as much as possible using forced alignment software tools.

There are many forced alignment software tools available, most of them focusing on phone alignment of English-language material (for a comprehensive list see Pettarin, 2018). Only a small subset of these support multiple languages and even the ones that

https://doi.org/10.14712/24646830.2025.20

do often do not include Czech as an option. Among the most used alignment tools with support for multiple languages generally employed in phonetic research are the Munich alignment system MAUS (Schiel, 1999) with its web-based implementation, and the Montreal forced alignment software (McAuliffe et al., 2017). Tools developed specifically for alignment of Czech-language material include Prague Labeller (Pollák et al., 2005, 2007) and more recently, Labtool (Patc et al., 2015) and Prak (Hanžl & Hanžlová, 2023). There is also a forced alignment tool directly integrated in Praat (Boersma & Weenink, 2023), which is useful as an easy-to-access tool. It has, however, potential for further upgrading and development (personal conversation with Paul Boersma, 2023). Some features of these tools will be elaborated on below.

When editing a sound and textgrid simultaneously in Praat's *View & Edit* window, selecting the menu item *Interval / Align Interval* (or Ctrl+D) will add a word and phone tier to the existing textgrid with time-aligned intervals. The option works for many languages (including Czech) and can align single words or short phrases. The alignment tool in Praat uses a speech synthesizer to create an audio track based on the provided orthography. The created sound is then compared to the provided audio and aligned using dynamic time warping (Boersma et al., 2023). This makes the alignment option simple to implement, but also limits its use when aligning longer sequences, especially ones containing pauses, as these are not reliably identified by the algorithm. The option also works exclusively from the *View & Edit* window and doesn't have a setting for automatic alignment of multiple files.

A very widely used forced alignment software with easy access and no installation is the Munich Automatic Segmentation System (MAUS, Schiel, 1999) with its web service (Kisler et al., 2017). The use of the MAUS web service is free for members of academic institutions for non-profit use, otherwise users must obtain a license to use it (Bavarian Archive for Speech Signals, 2018). The alignment software supports over 30 languages, with multiple dialect variants for English and German. The list of supported languages does not, however, include Czech. The web-based interface makes forced alignment using MAUS easily accessible with only a web browser and internet connection, but obscures the source code and does not therefore allow tweaking the way the system runs or implementation of the user's own models, such as models trained for other languages.

The forced alignment software that was until recently very commonly used to align Czech recordings for purposes of phonetic research is Prague Labeller (Pollák et al., 2005, 2007), developed at the Czech Technical University. The aligner uses HTK GMM models and employs a rule-based pronunciation generator. The software was a state-of-the-art forced alignment tool at the time of its development and was in consistent use for more than a decade at the Institute of Phonetics, Charles University. We have included a comparison of this aligner with our newly developed tool in Section 5. Development of new language model software options in the recent years has led to the tool becoming impractical to install with its dependencies. There is also a demand for higher accuracy of the automatic alignment. Additionally, the software is not freely distributed and runs only under Windows, which prevents the public, including students of phonetics, from using the alignment software with their own devices.

There has also been an effort to develop a newer forced alignment tool focused on Czech, implementing HMM-based phonetic segmentation using Kaldi instead of HTK

models (Patc et al., 2015). The main focus of this tool was to enable detailed study of pronunciation variation in spontaneous Czech speech through automated segmentation and variant detection, integrating Kaldi's acoustic modelling techniques. Experimental results showed that Kaldi-based models provided more consistent and precise phone boundaries compared to older HTK-based methods. Despite these results, the software is not in public distribution and has not been widely deployed by Czech phoneticians.

A similarly recent Kaldi-based forced alignment software with a wide range of supported languages, including Czech, is the Montreal Forced Aligner (MFA, McAuliffe et al., 2017), which is available under a MIT license (Opensource.org, 2025). The installation of the MFA requires Kaldi as a dependency and the download of models for the desired language. The aligner includes several pre-trained models for Czech, the more basic ones being trained on the same CommonVoice (Ardila et al., 2019) database as the original model in Prak (Hanžl & Hanžlová, 2023). More advanced models use training data from larger databases, including paid datasets. The option to choose from multiple models allows some customization for the user, nevertheless, the phone sets implemented in the models for Czech may not be sufficient for the purposes of detailed phonetic research (as discussed in Hanžl, 2023). The MFA tool does allow for implementation of the user's own models if so desired, so these issues can be resolved, but the installation is still quite convoluted, so developing a new easy-to-install tool along with more precise models is a logical step in the process of ameliorating Czech forced alignment options.

The most recent tool which specifically aims to provide a streamlined installation process as well as resolve known automatic segmentation issues and improve phone alignment of Czech recordings (with the option to train and implement models for other languages) is Prak (Hanžl & Hanžlová, 2023). Similarly to the MFA, this software is freely available on all major computer operating systems under a MIT license. However, unlike most forced alignment tools that are in wide use, Prak requires only minimal dependencies and aims to keep its architecture simple in order to not only be usable as a user-friendly alignment tool, but also to enable other researchers or programmers in the future to build on it without restrictions. The default model provided in the free distribution of Prak is trained on CommonVoice (CV) Czech recordings, as mentioned above. Due to its simplicity, the tool is useful even for non-phoneticians, such as students in a pronunciation class, to help navigation in longer sound files by quickly obtaining an overview of the contents of a recording.

While the aligner with this CV-trained model significantly outperforms the forced alignment softwares for Czech that have been in use before, the phone boundaries still often need to be moved manually after the automatic alignment to achieve the precision needed in phonetic research. The newest step in the development of Prak therefore was to train a new model in collaboration with the Institute of Phonetics, Charles University, which would use manually aligned recordings and HuBERT (Hsu et al., 2021) embeddings in addition to the original training dataset in order to more closely mimic human behavior based on established Czech segmentation rules (Machač & Skarnitzl, 2009), hopefully further reducing the amount of manual labor necessary after using the forced alignment software.

In this paper, we aim to present Prak and its functionality in a comprehensive way, as well as provide detail about the training of the new model and compare both Prak

models with the output from Prague Labeller as the long-standing predecessor in Czech phonetic alignment. Our goal is twofold: first, to introduce our software to its intended users; and second, to present the underlying concepts and development strategies in sufficient detail to enable future developers and researchers to build upon it. To that end, this paper is structured to reflect both the practical and methodological dimensions of the tool.

We provide an outline of the installation process, including software prerequisites and integration with the Praat environment. This is followed by a description of the pronunciation modeling framework, including built-in replacement rules, the exceptions file, and the modular Finite State Pronunciation Blocks. We then detail the training procedure for the HuBERT-based model, discussing both the use of additional manually aligned data and the architectural considerations that informed our design choices. Finally, we evaluate the performance of the system, comparing both Prak models with the Prague Labeller using manually aligned data as a reference, and report on alignment accuracy in terms of phone identity and boundary placement.

2. Design and installation

2.1 Installation of the software and prerequisites

The use of Prak requires only two external software tools, namely Python 3 (Van Rossum & Drake, 2009) with the PyTorch (Paszke et al., 2019) and TorchAudio (Yang et al., 2022) libraries. The installation of these prerequisites is clearly outlined in the official Prak installation instructions (Hanžl & Hanžlová, 2025), and no knowledge of programming or speech technology is necessary to complete the setup. The documentation provides step-by-step commands that can be run via command line, and offers platform-specific guidance where relevant. This makes the software accessible to phoneticians and linguists as well as other researchers or students who may not necessarily have a technical background. At the same time, this simple and modular structure ensures that the code remains easily readable and modifiable for programmers or developers who wish to extend its functionality.

The Prak installation process further involves only downloading the Prak code (via Github or as a zip file) and choosing the desired model for alignment. Available options include the basic Prak-CV model (as presented in Hanžl & Hanžlová, 2023) or the more fine-tuned, recent model based on HuBERT (Hsu et al., 2021). Details regarding the HuBERT-based model and its properties are provided in Section 4. Once installed, Prak can be run via command line or through a script which integrates Prak's functionality into the GUI in Praat (Boersma & Weenink, 2023) while also adding supplementary features, as described in the following section.

2.2 Praat GUI integration and additional functionality

Apart from functioning directly from the command line, Prak provides a Praat script which embeds the main Python forced alignment tool and can be added to Praat's dynam-

ic menu for easy access. The script also performs several assessments of the input files and provides additional options for the alignment. First, basic file checks are performed. The number of sound and textgrid files are counted and compared to ensure all desired recordings have a text input to be used in the alignment. The Praat interface also provides an additional option to use only one text input to align multiple sound files. Sound and textgrid names are also compared and in case of name mismatch, the user is prompted to decide whether the combination of the files with different names was deliberate. When working with multiple files, this check can be overruled and sounds aligned by the order in which the items are open in Praat.

The contents of each textgrid provided to the tool are also reviewed in order to determine the source text correctly. The tool expects the tier containing the source text to be named "phrase" and outputs three tiers after performing the alignment: a "phone" tier containing the phone boundaries, a "word" tier containing word boundaries and a "phrase" tier with the original source text. This is modelled after the Prague Labeller (Pollák et al., 2005, 2007) output, established as the standard at the Institute of Phonetics, Charles University. An example of the output textgrid is presented in Figure 1. The script integrating Prak into the Praat UI firstly checks that the source textgrid doesn't contain a non-empty "phone" tier to prevent accidental overwriting of files. If such a tier is found in the textgrid, the user is notified of this circumstance and can choose to either stop the script or continue and overwrite said file. Similarly, if a "phrase" tier is missing in the source textgrid, the user is prompted to identify the tier containing the source text which is to be used.

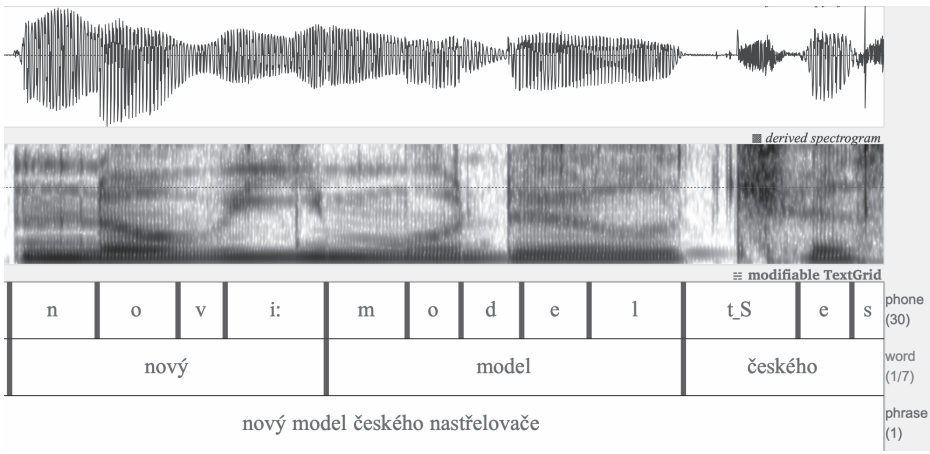


Figure 1 Example of an output textgrid after forced alignment using Prak.

If all assessments of source files are successful, the sounds and their corresponding text sources are fed to the Python tool to proceed with forced alignment. In the alignment process, textgrids containing the source text are replaced with output files containing the three tiers described above. All other tiers that may be in the original textgrids in addition to the source text are ignored and are not part of the final aligned files. All output textgrids are also renamed after the sound files they correspond to.

3. Pronunciation rules

The pronunciation generator in the original Prak used several novel principles, trying to remedy deficiencies of pronunciation generators used in previous decades, especially trying to make the set of rules manageable long term and the level of detail adjustable should the scientific research at hand require such a change. An overview of the phonetic alphabet used, text cleanup issues and basic design of the pronunciation generator was presented in our original Prak introduction publication (Hanžl & Hanžlová, 2023). While we mostly reused the generator unchanged for the new version of Prak, the available descriptions of the design principles and implementation are rather superficial, leaving direct inspection of the source code as the sole option for researchers seeking insight into the details. We therefore take the opportunity to describe these components in a format that is more accessible and comprehensible to readers.

The primary user group targeted by the design of Prak are researchers who are likely to fine-tune the pronunciation rules logic. Simple dictionary-based approaches often employed for English are largely insufficient for Czech, which is a highly flexible language, necessitating many entries for all the forms of every word and making manual ad-hoc additions of new words to the dictionary quite cumbersome. Usual approaches based on replacement rules are also hard to apply, mainly due to large consonant clusters with complex assimilation rules in Czech pronunciation, where, among other processes, regressive assimilation of voicing applies to all viable consonants within said cluster (Skarnitzl, 2011, p. 123). Coping with the voicing or devoicing of phones presents a considerable challenge. As a result, the scope of our pronunciation generator tool is rather narrow, addressing the specific needs of phoneticians working with the Czech language. Nonetheless, there is potential for reusing components of our software in other languages with similar phonotactic and assimilation patterns, such as Polish, Slovak, Russian, or even Armenian (Kuldanová et al., 2022; Pavlík, 2009).

After decades of experience with the replacement-rule-based Czech pronunciation generator used in Prague Labeller (Pollák et al., 2005, 2007), we decided to address the main known shortcoming: The rules table grew to hundreds of entries over years of use, and while this approach worked, inserting a new rule in the correct position among existing ones became a highly expert task, as it always required verification on a large corpus of previously generated pronunciations, identifying all changes caused by the new rule, and determining whether they served the intended purpose. Making pronunciation rules position-independent was therefore an important design goal of Prak. This initially seemed difficult, as the rule order also corresponded to gradually changing layers of representation which started with graphemes and gradually progressed through phones to allophones. However, we were able to find a practical solution, structuring the pronunciation processing in two layers using two different approaches:

1. A set of replacement rules without any human-defined order. The rule with the longest match is applied first. The rest of the word is then subject to more possible replacements but whichever part was already touched by another replacement rule is not affected by any other. This part can be easily used to specify pronunciation of "foreign looking" substrings and stay close to the graphemes.

2. A Finite State Machine based layer mainly taking care of Czech assimilations. This layer can be adapted to a particular research goal and the corresponding details of phonemic representation but does not require changes as new speech material is being added.

3.1 Built-in replacement rules and Exceptions file

Built-in rules deal with the most common patterns of foreign pronunciation in Czech. They can also serve as a didactic example for entries in the Exceptions file, as the format of both is the same and in practice, they are mixed by Prak into a single optimized structure with priority assigned by match length. Any built-in replacement rule can therefore be overridden in the Exceptions file simply by using a longer (more specific) string to be replaced. A notable feature of the replacement rule engine in Prak is its ability to consider multiple replacements. Each rule specifies a substring to be found in a word and lists one or more possible replacement strings. Selection of the right pronunciation version is later determined by the acoustic properties of the signal being processed.

As mentioned above, in addition to the built-in replacement rules, the Prak source contains an Exceptions file, where further pronunciation rules can be specified by the user. The built-in pronunciation generator is very meticulous in considering possible assimilations (of various kinds) and even glottal stop presence, which is not always required by the orthoepic norm (Volín, 2012; Volín & Skarnitzl, 2018, p. 22) and can have multiple acoustic realizations (Machač & Skarnitzl, 2009, pp. 125–131). However, due to the nature of the pronunciation generator, as described below, Prak has limited lexical knowledge and therefore may require additional input for handling cases that fall outside the scope of general pronunciation rules.

The Exceptions file is consulted when Prak is invoked from Praat. It uses very simple entries, modelled after the built-in rules, which are easy to follow and add to as the need arises. The file allows users to manually specify the pronunciation of strings at or below word-level which then override any other rule that may otherwise be applied to said strings within the default processing. This is particularly useful for dealing with proper names, loanwords, abbreviations, or unusual morphophonological irregularities that are difficult to capture systematically. Each entry in the exceptions file maps a written string directly to its target phonetic or allophonic representation, ensuring accurate alignment in contexts where automatic rule application could be unreliable. Specific instructions along with examples of added pronunciation rules can be found on the Prak installation page linked in Section 6.1.

3.2 Finite State Pronunciation Blocks

Finite State Transducers (FSTs) have appeared as a unifying concept in some speech recognition systems in the past. However, using backward-going FSTs to translate one symbol sequence into another as a pronunciation assimilation tool is a rather unique feature of Prak, and we believe it is a highly efficient option for Czech (and potentially for other structurally similar languages). Furthermore, we use non-deterministic FSTs, which may suggest multiple output symbol options. The convolution of several such

non-deterministic FSTs, together with the potentially multi-option replacement rules described above, creates a very powerful tool capable of handling complex Czech assimilations, even in words of foreign origin. At the same time, describing possible pronunciations remains mentally manageable, as it is decomposed into clearly understandable parts – rules handling foreign elements at a level close to graphemes, and FSTs dealing with assimilations, with each simple FST addressing just one phenomenon. The resulting pronunciation options can then be visualized as a so-called "sausage" structure, offering a more accessible alternative to Directed Acyclic Graphs.

The need for the use of FSTs arises from ongoing efforts to enumerate the possible assimilation changes in Czech consonant clusters. As mentioned in Section 3.1, in the Czech language, many of these changes are like a domino effect going backward in a sequence of possible phones. The change can be rather far reaching, and consonant clusters can be remarkably long. For example, in the approximately 6,000 different words in our training set, nearly 700 distinct consonant clusters were identified, 17 of which had a length of five or more characters. Capturing the essence of assimilation logic in FSTs turned out to be a practical solution to dealing with this vast variety.

Figure 2 demonstrates a FST handling backward assimilation of the voiced/unvoiced property. The current FST state depicts the influence being exerted on the phone to the left. For clarity, only a subset of the edges is shown in the figure (edge labels are in the format Input/Output). The word "kdyby" is processed right to left, changing the consonant cluster "kd" to [gd].

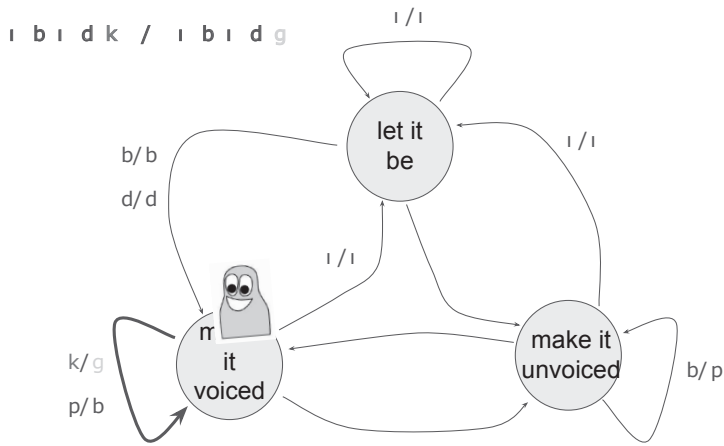


Figure 2 Illustration of regressive assimilation of voicing, as processed through backward traversal of the voiced/unvoiced property by a FST. Token (pawn) represents the current state.

The current state is depicted as a token (pawn), which can travel along edges, consuming an input symbol I and producing an output symbol O when moving along an edge labeled I/O. Should there be multiple possible edges with the same matching input symbol I, the token "clones" itself into multiple tokens, each following a different path and creating separate branches in the pronunciation "sausage" graph. When clones meet at the same state, the tokens merge again along with the pronunciation branches they represent.

A typical example of this branching occurs at word boundaries. Depending on the degree of word separation, the assimilation domino effect may either cross the boundary or stop at it. Another example – modeled by a different FST – would be pronunciation of the word "galantní", where the cluster "ntn" can be realized as [ntɲ], [ncɲ], or [ɲcɲ]. This represents another case of non-deterministic regressive assimilation that can be effectively modeled by FSTs. The actual Prak algorithm is somewhat more nuanced than simply applying non-deterministic replacement rules followed by a convolution of several backward-going non-deterministic FSTs, but a large portion of Czech pronunciation logic can indeed be captured using this scheme. Further details regarding the pronunciation generator and its implementation can be observed in the Prak source code.

4. Training the new HuBERT model

The original Prak release prioritized simplicity, easy maintainability of the codebase and simple installation. The neural network architecture of the Prak-CV model used only a very simple stack of ReLU (Rectifying Linear Unit) layers for phone classification. The goal was to design the system in a way that would be accessible to researchers who are mainly interested in phonetics, rather than artificial intelligence specialists who are more likely to experiment with complex and rapidly evolving architectures. Even under this restriction, the improvement in precision was substantial when compared to the previously employed Gaussian mean models from the HTK toolkit era. Nonetheless, transformer-based embeddings have become so widespread in recent years (Lin et al., 2022) that we decided to incorporate them in an improved version of our phone classification module.

Maintaining relative simplicity was still among our aims for the new model, therefore, we selected a freely available neural network model that computes HuBERT embeddings (Hsu et al., 2021). This network is pre-trained on a mix of many languages and coefficients of this neural network are available for automatic download from public servers at the time of first inference on a particular computer. Even though it would be possible to fine-tune coefficients of the HuBERT network, doing so would compromise the simplicity of installation and coefficients of the fine-tuned HuBERT network would need to be included in the distribution of Prak itself, making the distribution package many times bigger. Instead, we decided to use the HuBERT network in its default form and train an alternative of our original simple ReLU stack which would use HuBERT embeddings as an additional input. This approach proved to be largely sufficient for our goal of achieving more precise identification of phone identities and time boundaries.

4.1 Using additional manually time-aligned training data

As mentioned in the introduction, the original Prak phone classifier model was trained exclusively on the freely available CommonVoice dataset (Ardila et al., 2019). The goal was to enable easy portability to other languages – the CommonVoice database is available for many languages, not only Czech – as well as keep the prospect of future precision improvements with retraining on the ever growing CommonVoice data. All phone boundaries in

the original model were automatically derived during training, as the CV dataset contains only recordings of sentences and their orthographic transcripts. Any exact match between manually annotated and automatically derived phone boundaries is therefore just a result of a coincidental match between human judgment and the system's estimation of the most likely boundary location (nevertheless, this match is often notably close).

The match achieved by the original Prak model was still a significant improvement over the tools used previously – this fact being a tribute not only to Prak but also to the human team working on the reference labeling, trying to make evidence based decisions based on an agreed upon sensible set of rules. Convergence of the time boundaries derived by these two independent processes, one based on human expertise and manual annotation and one purely data driven, was remarkable. However, greater precision was still achievable, and the long-term goal remained to train a model which would place time boundaries where human researchers expect them, given their specific research approaches and objectives. In line with this goal, we added data with manually time-aligned boundaries to the training dataset for the new HuBERT-based Prak model.

The dataset used for the training of this model in addition to the CV recordings used in the original model was a corpus of manually corrected time-aligned recordings, generously provided by the Institute of Phonetics, Charles University. This additional dataset consists of 1435 recordings with a total duration of 5 hours and 15 minutes. All recordings in the dataset were aligned by a forced alignment software (the majority by Prague Labeller) and subsequently manually checked and adjusted to comply with the Czech standards for phonetic segmentation (Machač & Skarnitzl, 2009)[6]. The manual alignment was done for the purposes of conducting phonetic research (see for example the research by Volín & Skarnitzl, 2022, which uses a subset of this corpus) rather than training a model for forced alignment software. This corpus should therefore reflect the needs of Czech phoneticians in terms of the standards expected when conducting research. The exact contents of our time-aligned training dataset are presented in Table 1.

Table 1 Overview of the contents of the corpus of manually corrected time-aligned recordings used in addition to the CommonVoice dataset in training the new Prak model.

type of text	type of speaker	total duration	n female speakers	n male speakers
audiobooks	professional	134 m 10 s	6	5
poetry reading	amateur	89 m 22 s	14	12
radio news	professional	59 m 25 s	5	11
	amateur	31 m 49 s	8	1
total		314 m 46 s	33	29

The dataset consists of recordings from three different genres, recorded under different conditions by both professional and amateur speakers. The first genre, accounting for approximately 2.25 hours of the recordings, were spoken narratives (i.e. storytelling)

[6] Based on our approximation, the labelling of this corpus of recordings took more than 300 hours of manual labour in addition to using a forced alignment software.

extracted from audiobooks recorded by experienced actors in professional studios and produced by renowned publishers. The extracts used in the training dataset were read by 6 female and 5 male speakers.

The second type of recordings in our training dataset were readings of poetry, the total duration of which was around 1.5 hours. These samples of poetry reciting were recorded by 14 female and 12 male speakers. The recordings were done in the sound treated studio of the Institute of Phonetics in Prague. The speakers were volunteering students of philology with an interest in poetry.

The third genre included in the dataset were two types of news reading with a total duration around 1.5 hours, similarly to the poetry reciting samples. Two thirds of this subset (around 1 hour) consist of recordings of authentic news-bulletins from Czech radio broadcasts, recorded by 16 (5 female, 11 male) professional speakers. The remaining third of the samples are texts taken from said Czech radio broadcasts read by 9 volunteering students (i.e. nonprofessional speakers; 8 female, 1 male) in the same studio the poetry reciting was recorded.

4.2 Details of the HuBERT model

The HuBERT model computes a transformer-based embedding, similar in principle to word representations used in many translation systems and artificial intelligence dialogue systems, and analogous to amino-acid context-aware representations in modern approaches to analyzing or even synthesizing proteins in biochemistry, among other uses. The transformer-based processing has become a highly successful overarching paradigm (Lin et al., 2022; Vaswani et al., 2017). HuBERT, in particular, applies this paradigm to short (20 ms) segments of the speech signal, representing these as long vectors of numbers describing not only the spectral characteristics of the specific sound (as traditional cepstral features do) but also the meaning of the sound chunk in a particular context.

The HuBERT model is pre-trained on a large multilingual corpus of unlabeled speech. Similar to training procedures in other domains, this process involves masking short segments of the speech signal and requiring the model to reconstruct these masked parts as precisely as possible. This forces the model to capture long-range dependencies in the speech signal, enabling it to perform speaker adaptation and other phone-level analyses, ultimately acquiring enough knowledge to fill in the missing segments with high confidence (Boigne, 2021). In practical use of the pre-trained HuBERT model, our goal is often different, as we typically have the complete recording without any missing segments. We instead leverage the internal representation developed during the training and reuse it for particular tasks at hand. As it turns out, these internal representations are very rich and context-aware and can be applied to a variety of tasks with notable success.

There are two ways to use the pre-trained transformer-based models:

1. Keep the pre-trained model unchanged and train an additional neural network connected to it. The added network uses the pre-trained representations as inputs and is trained to produce the desired outputs.
2. Train not only the additional network but also fine-tune the pre-trained model itself.

We decided to employ the first approach, which provided satisfactory results for our purposes, so we kept it as the final solution. As mentioned above, while the second

approach could theoretically achieve even better results in our tests, it would also come with the technical disadvantage of having to distribute the modified HuBERT parameters together with Prak, making the installation package significantly larger. Apart from this technical nuisance, there is also another risk: further training of the HuBERT coefficients could actually reduce Prak's precision in practice due to overfitting (see Chicco, 2017), as our training datasets are quite small for a model of HuBERT's size. Given all these considerations, we opted to use the unchanged HuBERT coefficients.

There are multiple versions of the HuBERT model – the BASE version, pre-trained on 960 hours of unlabeled audio from the LibriSpeech dataset (Panayotov et al., 2015), and the larger LARGE and XLARGE models, pre-trained on 60,000 hours of speech. Given the relative simplicity of our phone alignment task – compared to other HuBERT applications such as speech recognition – and our overall goal of maintaining simplicity, we opted for the smallest BASE model (Torchaudio Contributors, 2024).

The original Prak used a 10 ms time resolution for its cepstral features, while HuBERT uses 20 ms time steps. We certainly did not want to make Prak-detected time boundaries more coarse-grained (on the contrary, we would even consider a 5 ms time step for future work), so we simply repeated each HuBERT output vector twice. This way, HuBERT contributes mainly the long-term contextual information, while the 10 ms cepstral features allow for a sharper local resolution.

The HuBERT model uses a multilayer transformer input stack, computing different embeddings at each level. Each additional layer of this stack should theoretically produce increasingly more abstract and more context-aware embeddings, making the top layer the most information rich. In practice, any layer can be used as input for the subsequent processing, not restrained to the top one. Using one of the lower layers offers the potential benefit of computational savings, as the upper layers do not have to be computed. We therefore trained several variants of the system, aiming to identify the lowest layer that still allows the system to operate with negligible precision loss compared to the best (likely the topmost) transformer stack layer being used. After multiple training attempts, we selected layer 7, though the differences in performance across layers were relatively minor.

HuBERT BASE uses a 16 kHz waveform as input (the same as the original Prak). Instead of MFCC, the pre-trained HuBERT model uses a 7-layer Convolutional Neural Network which does a learned feature extraction, generating a 512-dimensional vector every 20 ms. These vectors are then processed by a stack of 12 transformer layers, each using 12 attention heads. Each 20 ms unit is represented as a vector of 768 numbers when passing between layers. This is the representation that is forked to the Prak phone classifier network.

The original Prak model used 13-dimensional MFCC features with 9 context frames on both the left and right, resulting in $(9 + 1 + 9) \times 13 = 247$ numbers, further augmented by $4 \times 13 = 52$ speaker adaptation values, for a total of $247 + 52 = 299$ values every 10 ms. The HuBERT-based Prak model retains this entire representation and concatenates it with the 768-dimensional vectors from HuBERT's 7th transformer layer (each vector being used twice, as HuBERT operates with 20 ms chunks), resulting in $299 + 768 = 1067$ values every 10 ms. This representation is fed to the 3-layer ReLU stack with an internal vector size 100, followed by a final softmax layer to produce phone probabilities.

Among the phone probabilities is also the probability of a «silence» sound which includes not only a real silence but also all non-speech events like breaths or hesitations, as encountered in the training data. Common Viterbi decoder then decides globally optimal phone identities and silence boundaries, given the choices proposed by a pronunciation generator.

5. Evaluation and comparison of phone alignment

5.1 Method and material

In order to evaluate the performance of our new HuBERT-based model, we compared the output of Prak alignment using this model with both the previous Prak-CV model (Hanžl & Hanžlová, 2023) and the Prague Labeller (Pollák et al., 2005, 2007), using manually aligned recordings provided by the Institute of Phonetics as the ground-truth evaluation baseline. We measured the percentage of phone identity mismatches compared to the baseline, as well as the percentage of phone boundary misalignments.

We contrasted the generated pronunciations, counting phone insertions, deletions and substitutions. Direct comparison of phone identities determined by the forced alignment tools is not straightforward, as the phone sets differ based on the inventories used by each aligner. For instance, compared to the Prague Labeller, Prak also detects glottal stops, distinguishes between voiced and voiceless "ř" or accounts for assimilations at word boundaries. Manually time-aligned textgrids may, on the other hand, reflect slightly different approaches to phone identity labeling, depending on the research questions for which the recordings were originally intended. Additional annotations of some segments, such as syllabic [l̩] or [r̩], may also be present and contribute to the phone identity mismatch percentage.

At places where phone identity matched, we measured time shifts of the phone boundaries relative to the manual reference. While the phone identity may in some cases be subject to interpretation, as discussed above, the time positions in the manually aligned reference data should be in compliance with the segmentation standard for Czech (Machač & Skarnitzl, 2009) and can therefore be considered accurate for the purposes of Czech phonetic research. An important parameter affecting the efficiency of manual correction of automatically aligned files is the frequency of boundary misalignments that require adjusting multiple phone boundaries to correct the error. We therefore used the number of boundary shifts exceeding thresholds of 100 and 200 ms as a quality measure of major boundary misplacement.

For our evaluation, we used a subset of the phonetic corpus presented in Section 4.1. We selected 156 recordings (balanced across the file subtypes in the dataset) that were not used at any stage of training for our HuBERT-based model or any other model. These recordings have a total duration of approximately 30 minutes and contain about 20,000 individual phones. We obtained the original output TextGrids of forced alignment produced by the Prague Labeller without any manual correction, and we performed forced alignment of the same recordings using Prak with both the CV and HuBERT models. The source text was copied from the manually labeled files and used as input in all three alignment iterations.

When we examined individual cases of divergence between manual labeling and Prak-HuBERT labeling, we found that most differences were caused by non-uniform manual labeling, for example, the use of different phone sets, the omission of certain phenomena, or, conversely, the inclusion of additional detail. While we invested substantial effort in automatic normalization of all the manually labeled data to a common standard, this process had inherent limitations. To gain additional insight, we further double-checked the manual labeling in our test set for errors and compliance with the selected transcription method in cases where it diverged from the Prak-HuBERT labelling. We then re-evaluated the Prak-HuBERT model and observed, for instance, nearly ten times fewer boundary shifts exceeding 0.1 s. We therefore added an additional test for shifts over 50 ms to gain finer granularity. While the results in this additional table are truly impressive, they are no longer strictly objective and should be interpreted as suggesting that Prak-HuBERT errors are approaching the limits of what can be reliably measured.

5.2 Results and discussion

Table 2 shows the percentage of errors in the phone identity mismatch and boundary misplacement tests, as well as the cumulative results of these tests, comparing the output of Prak's models and the Prague Labeller with data from manually time-aligned files. It is evident that Prak generally outperforms its predecessor in all types of tests provided, indicating a significant decrease in errors leading to manual corrections requiring adjustment of more than one boundary.

Table 2 Percentage of phone mismatch and boundary misplacement, comparing the output of Prak's two models and the previously most used forced alignment tool with manually time-corrected recordings.

test type	Prague Labeller	Prak-CV	Prak-HuBERT
phone identity mismatch	6.61%	1.88%	1.12%
match, but misplace ≧ 0.1 s	4.28%	0.36%	0.04%
match, but misplace ≧ 0.2 s	3.22%	0.09%	0.00%
mismatch or misplace ≧ 0.1 s	10.89%	2.24%	1.16%

An improvement can also be observed between the two Prak models, with the HuBERT-based model reducing boundary misplacement errors of 100 ms or more from 0.36% in the Prak-CV model to 0.04%. The new model also virtually eliminates errors involving boundary misplacements of 200 ms or more, which suggests that word-level identification by this model is highly reliable.

Prak-CV and Prak-HuBERT use the same pronunciation module. However, this module often generates variant pronunciations, and Prak-HuBERT makes better use of the acoustic evidence to select the correct variant. This explains the improved performance of Prak-HuBERT, even in terms of phone identification. The phone identity mismatch could be further reduced by using the pronunciation exceptions file. This is an expected practice when using Prak for research purposes, however, we did not generate a dedicated exceptions file for our test set.

Results of additional testing using the Prak-HuBERT model with test data modified to comply with our selected labeling method are shown in Table 3. Compared to the initial test, boundary misplacements of 100 ms or more were reduced ten times, and phone identity mismatch decreased by 0.3 percent following this adjustment. The additional test for boundary misplacements of 50 ms or more can be interpreted as capturing all major shifts, including ones smaller than the typical duration of one phone.

Table 3 Percentage of phone mismatch and boundary misplacement, comparing the output of Prak's HuBERT model with manually time-corrected recordings further double checked for errors and compliance with the selected transcription method.

test type	Prak-HuBERT
phone identity mismatch	0.946%
match, but misplace ≧ 0.05 s	0.133%
match, but misplace ≧ 0.1 s	0.005%
match, but misplace ≧ 0.2 s	0.000%
mismatch or misplace ≧ 0.1 s	0.951%

Overall, the results demonstrate that the HuBERT-based version of Prak significantly improves alignment accuracy over both the earlier Prak-CV model and the Prague Labeller. Reductions in phone identity mismatches and major boundary misplacements indicate a strong alignment with manually annotated data, while additional tests suggest that remaining errors are minimal and approach the limits of what can be reliably measured. These findings support the practical applicability of Prak-HuBERT in precise phone labelling for phonetic research in Czech.

6. Additional remarks

6.1 Public availability

We have decided to make the improved version of Prak, incorporating HuBERT and fine-tuning on additional time-aligned data, freely available for any purpose, with the only added requirement being the citation of relevant publications. The user of Prak now has two options:

1. Use the original model trained on CommonVoice, in which case Prak is available under the very permissive MIT license.
2. Opt for increased precision in phone boundary alignment, closer to practices established by the Institute of Phonetics, Charles University in Prague. In this case, the only additional requirement is that any publication benefiting from the improved model should cite the relevant publications of the Institute of Phonetics, as stated on the license page at time of installation.

Download of the software, along with usage details for both models, is available through the project website on GitHub: https://github.com/vaclavhanzl/prak

6.2 Future work

Prak was designed for the alignment of relatively short recordings, typically around one minute in length. Some internal algorithms (such as the search for the best alignment path and attention evaluation in the transformer layers) have roughly quadratic complexity, which makes processing slow for exceedingly long inputs. Manually splitting recordings into smaller parts is a simple workaround, but automating this step would be a much more user-friendly solution. This will require some additional research, as both the audio and the corresponding text must be divided into corresponding chunks, and the chunks must be large enough to preserve the contextual benefits provided by the transformer-derived embeddings. Nevertheless, such a feature would undoubtedly be welcome by users.

Additional testing of Prak's current and potential future models can also be conducted, for example examining the relevance of phone boundary precision in common phonetic measurements such as formant detection. Outputs based on measurement from purely automatically aligned data could be compared with results obtained using manually corrected time boundaries, providing insight into the level of precision required for reliable acoustic measurements in phonetic research.

Acknowledgments

We would like to thank the Institute of Phonetics, Charles University, for providing the training dataset of manually time-aligned recordings. Without the tedious work of all who have contributed to the corpus, training a model in order to reduce the amount of labor included in phone alignment for the purposes of phonetic research would not be possible.

REFERENCES

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., & Weber, G. (2019). Common Voice: A Massively-multilingual speech corpus. *arXiv Preprint arXiv:1912.06670*.

Bavarian Archive for Speech Signals. (2018). *Terms of Usage*. Version 4.0. BAS Web Services. https://clarin.phonetik.uni-muenchen.de/BASWebServices/help/termsOfUsage

Boersma, P., & Weenink, D. (2023). *Praat: Doing phonetics by computer.* [Computer program]. Version 6.3.14. http://www.praat.org

Boersma, P., Weenink, D., & collaborators. (2023). *Praat: Doing phonetics by computer* [Source code]. https://github.com/praat/praat

Boigne, J. (2021). *HuBERT: How to Apply BERT to Speech, Visually Explained*. https://jonathanbgn.com/2021/10/30/hubert-visually-explained.html

Chicco, D. (2017). Ten quick tips for machine learning in computational biology. *BioData Mining*, *10*(1), 35. https://doi.org/10.1186/s13040-017-0155-3

Hanžl, V. (2023). *Details of the Montreal FA*. Prak Wiki. https://github.com/vaclavhanzl/prak/wiki/Details-of-the-Montreal-FA

Hanžl, V., & Hanžlová, A. (2023). Prak: An automatic phonetic alignment tool for Czech. In R. Skarnitzl & J. Volín (eds.), *Proceedings of the 20th International Congress of Phonetic Sciences* (pp. 3121–3125). Guarant International.

Hanžl, V., & Hanžlová, A. (2025). *prak: Czech phonetic alignment tool*. https://github.com/vaclavhanzl/prak

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 3451–3460. https://doi.org/10.1109/TASLP.2021.3122291

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

Kuldanová, P., Hebal-Jezierska, M., & Petráš, P. (2022). *Orthoepy of West Slavonic Languages (Czech, Slovak and Polish)*. Ostravská univerzita. https://doi.org/10.15452/Ortoepieen.2022

Lin, T., Wang, Y., Liu, X., & Qiu, X. (2022). A survey of transformers. *AI Open*, *3*, 111–132. https://doi.org/10.1016/j.aiopen.2022.10.001

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

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. *Proceedings of Interspeech 2017*, 498–502. https://doi.org/10.21437/Interspeech.2017-1386

Opensource.org. (2025). *The MIT License*. Open Source Initiative. https://opensource.org/licenses/MIT

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5206–5210. https://doi.org/10.1109/ICASSP.2015.7178964

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An Imperative style, high-performance deep learning library. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 8026–8037.

Patc, Z., Mizera, P., & Pollak, P. (2015). Phonetic segmentation using KALDI and reduced pronunciation detection in causal Czech speech. *Text, Speech, and Dialogue*, 433–441.

Pavlík, R. (2009). A Typology of assimilations. *SKASE Journal of Theoretical Linguistics*, *6*(1), 2–26.

Pettarin, A. (2018). *A collection of links and notes on forced alignment tools*. https://github.com/pettarin/forced-alignment-tools

Pollák, P., Volín, J., & Skarnitzl, R. (2005). Influence of HMM's parameters on the accuracy of phone segmentation–evaluation baseline. *Proceedings of the 16th Conference Joined with the 15th Czech-German Workshop 'Speech Processing'*, *1*, 302–309.

Pollák, P., Volín, J., & Skarnitzl, R. (2007). HMM-based phonetic segmentation in Praat environment. *The XII International Conference Speech and Computer – SPECOM*, 537–541.

Schiel, F. (1999). Automatic phonetic transcription of non-prompted speech. *Proceedings of the XIVth International Congress of Phonetic Sciences*, 607–610.

Skarnitzl, R. (2011). *Znělostní kontrast nejen v češtině*. Epocha.

Torchaudio Contributors. (2024). *HUBERT_BASE*. https://docs.pytorch.org/audio/2.4.0/generated/torchaudio.pipelines.HUBERT_BASE.html

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Volín, J. (2012). Jak se v Čechách 'rázuje'. *Naše řeč*, *95*(1), 51–54.

Volín, J., & Skarnitzl, R. (2018). *Segmentální plán češtiny*. Univerzita Karlova, Filozofická fakulta.

Volín, J., & Skarnitzl, R. (2022). The impact of prosodic position on post-stress rise in three genres of Czech. *Speech Prosody 2022*, 505–509. https://doi.org/10.21437/SpeechProsody.2022-103

Yang, Y.-Y., Hira, M., Ni, Z., Chourdia, A., Astafurov, A., Chen, C., Yeh, C.-F., Puhrsch, C., Pollack, D., Genzel, D., Greenberg, D., Yang, E. Z., Lian, J., Mahadeokar, J., Hwang, J., Chen, J., Goldsborough, P., Roy, P., Narenthiran, S., … Shi, Y. (2022). TorchAudio: Building blocks for audio and speech processing. *ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, 6982–6986.

RESUMÉ

Článek představuje nástroj Prak, určený pro automatické časové zarovnání hlásek v češtině, který klade důraz na transparentní modulární strukturu a fonetickou přesnost. Kromě výslovnostního modulu pracujícího s pravidly a seznamem výslovnostních výjimek zavádí Prak nové využití nedeterministických, zpětně postupujících konečných překladových automatů (FST), zejména pro modelování regresivní asimilace v konsonantických shlucích. Dalším inovativním prvkem je integrace modelu HuBERT a trénování na rozsáhlém korpusu manuálně časově zarovnaných nahrávek, čímž se zvyšuje přesnost klasifikace hlásek, aniž by byla ovlivněna náročnost instalace a použití nástroje. Porovnání časového zarovnání hlásek s testovacím korpusem manuálně segmentovaných nahrávek ukázalo, že rozšířený model je výrazně přesnější v porovnání s předchozím Prak-CV modelem i dřívějším dlouhodobě používaným nástrojem pro časové zarovnání hlásek. Nový model výrazně snižuje pravděpodobnost výskytu hrubých chyb v určení hranic i nesouladů v identifikaci hlásek, čímž se úroveň zarovnání přibližuje standardům manuální segmentace. Nástroj je určen nejen fonetikům zabývajícím se češtinou, ale i vývojářům pracujícím s jazyky s podobnou strukturou.

Adléta Hanžlová
Institute of Phonetics
Faculty of Arts, Charles University
Prague, Czech Republic
adleta.hanzlova@ff.cuni.cz

Václav Hanžl
Department of Informatics and Chemistry, Department of Biochemistry and Microbiology
University of Chemistry and Technology Prague
Prague, Czech Republic

2025 ACTA UNIVERSITATIS CAROLINAE PAG. 61–83
PHILOLOGICA 3

PAUSING AND TEMPO VARIATION AS STRATEGIES IN SIGNALLING POETIC STRUCTURE

PAVEL ŠTURM

ABSTRACT
This study investigates how prosodic features reflect information structure and poetic organization during oral poetry performance. Specifically, we examined how repetition and structural position influence articulation rate (AR) and pause duration in spoken verse. Thirty-two native Czech speakers read three structurally comparable poems aloud, each differing in the presence and distribution of repeated lines. Poem 1 served as a baseline, containing no repetition; Poem 2 included a fully repeated final stanza; and Poem 3 featured repeated distichs within each stanza. Results showed that repeated lines (given information) were delivered at faster and more consistent rates than non-repeated lines (new information). Across poems with repetition, a gradual tempo decline followed by a tempo reset was observed, suggesting a strategic use of tempo modulation to signal textual recurrence. Additionally, pause duration reliably marked structural boundaries, with the longest pauses at stanza breaks. Discrepancies between syllabic and phonemic AR further highlighted the influence of phonotactic variability. Overall, the findings demonstrate that speakers intentionally manipulate prosodic timing to convey both informational and structural cues, enhancing listener comprehension of poetic form.

Keywords: poetry; information structure; phrasing; articulation rate; pauses

1. Introduction

Duration is one of the basic components of sound, employed in various ways for linguistic purposes. For example, it can support distinctions in vowel and consonant length, mark prominence, or cue prosodic boundaries within spoken utterances. Temporal aspects of speech may be expressed in terms of *duration* – such as the duration of individual phones or syllables – or in terms of *tempo*, defined as the rate of linguistic units over time. In this sense, the commonly observed phenomenon of final lengthening at the end of words or phrases may be more precisely characterized as gradual final *deceleration*, wherein speakers reduce their tempo to mark prosodic boundaries. Similar patterns can be observed in other acoustic parameters, including fundamental frequency (F0), intensity, and measures of voice quality (see Volín, Šturm, Skarnitzl, & Bořil, 2024, for evidence

https://doi.org/10.14712/24646830.2025.21

from Czech prosody). In addition, the presence of pauses in speech and their duration are also linguistically relevant parameters with a potential to signal the strength of prosodic boundaries (e.g., Zellner, 1994; Werner, Trouvain, & Möbius, 2022; Šturm & Volín, 2023).

There is considerable variation in pausing and tempo both across different speakers and within the speech of individual speakers. While speakers may display habitual temporal patterns, they also adapt their use of tempo in response to communicative intent, task demands, or genre conventions. For instance, Veroňková-Janíková (2004) found that the same speakers modified their overall tempo between read and non-read speech and varied their delivery between fairytales and other types of narration. In a similar vein, Volín (2022), analysing a larger sample of 24 speakers reading the same selection of texts, identified consistent tempo differences between two genres: news reading and poetry recitation. Specifically, speakers used a faster pace and exhibited greater tempo variation when reading the news than when performing poems.

Further evidence of systematic temporal variation comes from a large-scale study of spontaneous Dutch involving 80 speakers (Quené, 2005). The study examined both between-speaker factors (dialect region, sex, and age) and within-speaker factors (phrase length and the position of an utterance within an interview). While the topic of conversation was broadly controlled across participants, speech tempo, measured as average syllable duration (ASD) within inter-pause units, demonstrated complex patterns. Although initial models revealed significant differences across demographic groups, these effects disappeared once phrase length was considered as well. Longer inter-pause units were produced more rapidly, suggesting that speakers compress speech tempo over extended phrases (see also Crystal & House, 1990, for English).

Despite these findings, a considerable portion of tempo variation in Quené's study remained unexplained by demographic or structural predictors. This suggests the presence of systematic, communicatively driven tempo modulations within individual speakers. As Volín (2022) argues, such patterns likely reflect an underlying prosodic or temporal 'grammar' that governs how speech unfolds over time. Speech genres, speech styles, tasks, or units such as prosodic phrases may carry distinct norms of temporal realization shared among competent speakers.

Similarly, Nooteboom and Eefting (1994) emphasize the role of contextual factors in determining speech tempo. Their experiment found that ASD correlated well with the average number of phones per syllable in context-free sentences (thus replicating the above findings), but much less so – in fact, very poorly – in contextually embedded sentences. Phrase length may thus affect tempo primarily in cases involving simple, decontextualized utterances. The authors proposed that key factors include a phrase's position within a paragraph or the communicative relevance of its content (namely, given vs. new information).

The latter issue has been investigated by several researchers. For instance, Lieberman (1963) found that predictable words, which listeners could easily infer when omitted from the recording based on contextual cues, were pronounced faster and with greater acoustic reduction than less predictable words. This suggests that both speakers and listeners make use of the semantic and grammatical information included in meaningful utterances. Similar results were reported by Fowler and Housum (1987). In the production experiments, repeated words were shortened compared to their initial mentions in

monologues. Perceptually, listeners were able to distinguish between new and repeated words, with new words being more intelligible in isolation. Crucially, listeners used this information to integrate words into context.

The reduction of words with low information value can be viewed as an intentional process that facilitates comprehension of an utterance's informational structure (cf. Chafe, 1974). In this argument, important elements of speech (new, less predictable information) may be highlighted not only through melodic or energetic accentuation but also through localized tempo decreases. In contrast, redundant words may undergo various reduction processes, including temporal reduction (i.e., faster tempo).

A more fine-grained analysis was conducted by Eefting (1991), whose experiment investigated the effects of *accentuation* (focus) and *information value* (given vs. new information) on target word durations. While accentuation had a major influence, the durational effects of information value were in comparison minor and statistically insignificant, yet directionally consistent with expectations. These results suggest that information value alone has negligible durational consequences, although it may exert an indirect influence through its association with focus, as new information tends to be accented and thus lengthened (slowed down for processing). Eefting also cautioned that conversational speech might yield different results from her controlled, read materials.

The present study has several objectives. Primarily, it aims to demonstrate the value of investigating a specific and relatively underexplored genre: poetry reciting. A sample of speakers read/performed a collection of poems (see Volín, 2022), providing rich material for multi-level analysis. This paper will be limited to two specific research questions.

First, it explores how information structure affects pronunciation, using textual repetition as a proxy for given information.[7] Three poems were selected for this purpose: one serving as a control with no repetition, one featuring a repeated stanza, and another with two repeated lines in each stanza. We hypothesize that repeated passages will be spoken at a faster tempo than non-repeated passages or first mentions (cf. Fowler & Housum, 1987; Eefting, 1991).

Second, given the formal nature of poetry, we investigate whether its textual structure – visible in stanza, distich, and verse line layout – affects performance in measurable ways. Šturm and Volín (2023) examined four poems in relation to pausing: both pause frequency and duration increased at stanza ends. The present study seeks to replicate these findings on a different selection of poems and extend the analysis to speech tempo. We predict that pauses will be longer at the end of a distich and even longer at stanza boundaries. Furthermore, speech tempo is expected to decrease in corresponding verse lines, based on the known function of final deceleration in prosodic phrases (Paschen,

[7] This heuristic is limited, however, as textual repetition does not always map neatly onto givenness. In utterances at the so-called 'second instance level', where all elements are context-dependent, but one is highlighted prosodically (Firbas, 1979: 46; Svoboda, 1981: 4), the repeated material may still bear heavy ad-hoc contrast and therefore not function as straightforwardly 'given'. For example:
The meeting was successful. John finished the graphs and Peter secured the flight.
– So Peter came to the meeting?
– JOHN came to the meeting.
Here, the prosodically marked word *John*, although repeated, carries the highest communicative dynamism and becomes the most informative element in the final sentence.

Fuchs, & Seifart, 2022; Volín et al., 2024). In this way, stanzas may function similarly to paragraphs in prose, serving as higher-level organizational units.

2. Method

2.1 Material

The material analyzed in this study was drawn from a large corpus of poetry recitations (32 speakers, 60 poems) recorded at the Institute of Phonetics, Charles University in Prague. The description of the recording conditions and procedures is summarized in the next section (for more details, see Volín, 2022). Three poems (hereafter P1–3) were selected specifically for this analysis (note that they are different from those analyzed in Volín, 2022).

The selected poems were matched in overall length and structural features. Each consists of four stanzas, with four verse lines per stanza. The verse lines were comparable in length: all 11 syllables in P3, and alternating between 11 and 10 syllables in P1 and P2. All poems follow a regular rhyme scheme (either *abab* or *aabb*), although they differ in metre (P1 is dactylic, P2 iambic, and P3 trochaic) and in word count (106, 97, and 87 words, respectively).

A salient feature of all three poems is the internal structure of the stanzas, which are composed of two distichs (two-line units). These typically end with a full stop, while the first line of each distich is typically unpunctuated or ends with a comma. Crucially, each distich forms a coherent syntactic and semantic unit: the two lines belong together as a complete utterance.

The poems thus follow a consistent structure, repeated across all four stanzas (S1–4), with each stanza comprising four verse lines (VL1–4) organized into two distichs (D1–2). In subsequent analyses, three distinct positional types of verse lines will be considered: T1 = VL1 + VL3 (distich-initial lines); T2 = VL2 (distich-final but not stanza-final lines); T3 = VL4 (both distich- and stanza-final lines). An example from poem P2 is provided below:

S1	VL1	D1	T1	*Ty tóny duši rozrývají maně*	(11 syllables, 23 phonemes)
S1	VL2	D1	T2	*a pohádka to promrskaná dost.*	(10 syllables, 24 phonemes)
S1	VL3	D2	T1	*– Já o jedné jen snával karavaně*	(11 syllables, 25 phonemes)
S1	VL4	D2	T3	*a na poušti se bělá její kost –!*	(10 syllables, 23 phonemes)

It should also be noted that the number of phonemes per line varies independently of the number of syllables. In this particular stanza from P2, the phoneme count ranges from 23 to 25, with the entire poem exhibiting a range between 23 and 29 phonemes per line.

The primary criterion for selecting these three poems was the presence or absence of textual repetition, which varied systematically across the set. P1 served as a baseline, containing no repeated verse lines; each line in the poem was unique. P2 exemplified a stanza-level repetition, with the final stanza (S4) being a verbatim repetition of the initial stanza (S1), while the intervening stanzas differed. In contrast, P3 featured distich-level repetition: within each stanza, the second distich (D2) was identical across all four stanzas. The text of the poems is included in the Appendix.

2.2 Speakers

The material consists of poetry recitations by 32 Czech speakers (16 male, 16 female, mean age = 24.3 years, range = 19–33 years), all current or former philology students at Charles University with non-professional but relevant recitation experience. The participants were unaware of the study's purpose, as recordings originated from a student speech performance database. Each had sufficient time to prepare and practice, reducing errors. Crucially, participants were instructed to *recite* – not just *read* – the poems, treating them as expressive performances rather than neutral readings. This approach emphasized the aesthetic function of poetry. For full details of the recording procedure, see Volín (2022).

2.3 Measures

Pause duration

Delimitating pauses in speech is not a straightforward task. While many studies adopt fixed cut-off thresholds to define pauses, alternative approaches have challenged this practice (Werner et al., 2022; Šturm & Volín, 2023). These authors argue that imposing arbitrary thresholds can distort the natural distribution of pauses by systematically excluding shorter ones (cf. Campione & Véronis, 2002).

In this study, we adopted a threshold-free approach. Silent intervals that are intrinsic to speech sounds – particularly the closure phases of word-initial plosives or affricates – were annotated as part of the corresponding segment. The duration of word-initial plosives was generally constrained to a range of 50–100 ms, unless produced with marked emphasis on the word. Any remaining silent or filled interval preceding this annotation was considered a pause, regardless of its duration.

Articulation rate

Speech tempo can be quantified in various ways, depending on the domain of measurement, the treatment of pauses, and the choice of unit. In our material, the domain was self-evident: the verse line, which often – but not always – coincides with major prosodic phrases. As a result, local fluctuations in tempo within verse lines were not modelled. A more complex decision concerned whether to include pauses in the calculation of tempo. We chose to measure articulation rate (AR) rather than speech rate (SR) since our objective is to describe strategies used to signal poetic structure. When a speaker inserts a pause within a verse line without altering the speed of articulatory movements, SR changes significantly, while AR remains stable. Although such pauses may be linguistically meaningful, they are less clearly interpretable as cues to stanza structure, which is our key concern here. Moreover, pauses occurring at verse line boundaries – potentially relevant to stanza organization – were analyzed separately and therefore need not be absorbed into SR calculations.

Tempo measures also vary depending on the unit being counted: words, syllables, phones, or phonemes. On the one hand, Trouvain et al. (2001) showed that word-based measures are poorly suited to express speech tempo, identifying realized phone rate as the most reliable predictor of domain duration. On the other hand, from the perspective of

perceived tempo, syllables may be more informative. Pfitzinger (1998) conducted a perceptual study in which listeners ranked short speech segments by tempo and estimated their relative distances. Correlations with measured syllable and phone rates revealed that syllables aligned more closely with perceptual judgments. The strongest correlations were found for a linear combination of syllable and phone rates, with syllables weighted more heavily.

The complex syllable structure of Czech (Šturm & Bičan, 2021) also plays an important role. Verse lines with comparable syllabic ARs may differ substantially in phonemic AR. However, because the phonemic content of the text is determined by the poet's lexical choices, and our focus is on how speakers interpret and perform a fixed text, syllabic AR is prioritized in the analysis (but phonemic AR is also reported).

2.4 Analysis

The analysis of speech tempo was based on a total of 1536 tokens, as AR was measured for each verse line (3 poems × 16 verse lines × 32 speakers). In contrast, the data for pause analysis included fewer tokens (n = 1405). This reduction resulted from two factors: first, pauses were not measured after the final verse line of each poem; second, there were 35 additional instances in which speakers did not produce a pause following a verse line.

Two key variables were considered for each poem. In P1 and P2, verse lines (VLs) were categorized into three levels of STRUCTURAL TYPE (with treatment coding):

- T1: Distich-initial lines (VL1 and VL3)
- T2: Distich-final but not stanza-final lines (VL2)
- T3: Distich- and stanza-final lines (VL4)

In P3, the two T1 verse lines in each stanza required further differentiation, since one but not the other was repeated. As a result, a four-level VERSE LINE factor (VL1–VL4) was used instead of the STRUCTURAL TYPE classification applied in P1 and P2. The second factor considered across all poems was STANZA (S1–S4, treatment coded).

For each poem and parameter, a linear mixed-effects (LME) model was fitted using *R* version 4.2.1 (R Core Team, 2022) and the *lme4* package version 1.1.3 (Bates, Maechler, Bolker & Walker, 2015). Table 1 presents the six resulting models, including the specified effects and interactions. STANZA was modelled as a fixed effect in interaction with either STRUCTURAL TYPE or VERSE LINE, depending on the poem. For P1 and P2, random intercepts were specified for SPEAKER (32 levels) and ITEM (16 levels, corresponding to individual VLs); for P3, only speaker was included as a random intercept. Additional random slopes beyond those reported in the table could not be estimated due to convergence issues or singular fits.

In contrast to AR, which was modelled directly, pause duration was log-transformed prior to statistical analysis (cf. Šturm & Volín, 2023) and subsequently back-transformed for reporting and visualization. Tukey post-hoc tests were conducted using the *emmeans* package version 1.8.2 (Lenth, 2022), typically to compare STRUCTURAL TYPE within STANZA, and vice versa. *P*-values were adjusted by the Bonferroni method, based on the number of comparisons performed. The significance level was set at $\alpha = 0.05$.

Table 1 Specification of LME models for each poem (P1–P3) and parameter.

Poem	Parameter	Fixed effects	Random effects
P1	log(pause duration)	structural type * stanza	(1|speaker) + (1|item)
P2	log(pause duration)	structural type * stanza	(1|speaker) + (1|item)
P3	log(pause duration)	verse line * stanza	(1|speaker)
P1	AR in syllables/s	structural type * stanza	(1+structural type|speaker) + (1|item)
P2	AR in syllables/s	structural type * stanza	(1+structural type|speaker) + (1|item)
P3	AR in syllables/s	verse line * stanza	(1|speaker)
P1	AR in phonemes/s	structural type * stanza	(1+stanza|speaker) + (1|item)
P2	AR in phonemes/s	structural type * stanza	(1+structural type|speaker) + (1|item)
P3	AR in phonemes/s	verse line * stanza	(1|speaker)

3. Results

3.1 Pause duration

Figure 1 displays the raw (non-log-transformed) durations of pauses occurring after individual verse lines, excluding the final verse line in each poem. As expected, the distribution of pause durations is skewed toward shorter values, with a number of outliers at the upper end. Across all three poems, a clear effect of structural type emerged: pauses following ends of higher units (= T2, T3) were longer than pauses following VLs that were not distich-final (= T1). In most cases, pauses in T3 contexts were also longer than those in T2, reflecting the additional boundary at the stanza level. By contrast, no obvious effect of stanza appeared, as pauses seemed to have similar durations throughout the poem. Importantly, there was also no evident effect of repetition (in S4 of P2, or repeated distichs of P3).

No significant interaction was found between STRUCTURAL TYPE and STANZA in P1 ($\chi^2(5) = 3.5$, $p = 0.617$). While the inclusion of STRUCTURAL TYPE significantly improved the model ($\chi^2(2) = 22.2$, $p < 0.001$), STANZA did not contribute significantly ($\chi^2(3) = 2.0$, $p = 0.569$). Figure 2 on the left plots the predicted values from the LME model without interaction for the three levels of STRUCTURAL TYPE (values back-transformed to seconds). Pauses following the distichs (T2, T3) were longer than pauses in the middle of the distichs (T1), while stanza-final pauses (T3) were in addition longer than T2. All pairwise differences were statistically significant, as confirmed by Tukey post-hoc comparisons (see Tab. 2, top).

A nearly identical pattern was observed for P2 (Fig. 2 on the right, Tab. 2, bottom). Again, there was no significant interaction between STRUCTURAL TYPE and STANZA ($\chi^2(5) = 0.6$, $p = 0.989$), and only structural type emerged as a significant predictor ($\chi^2(2) = 23.1$, $p < 0.001$).

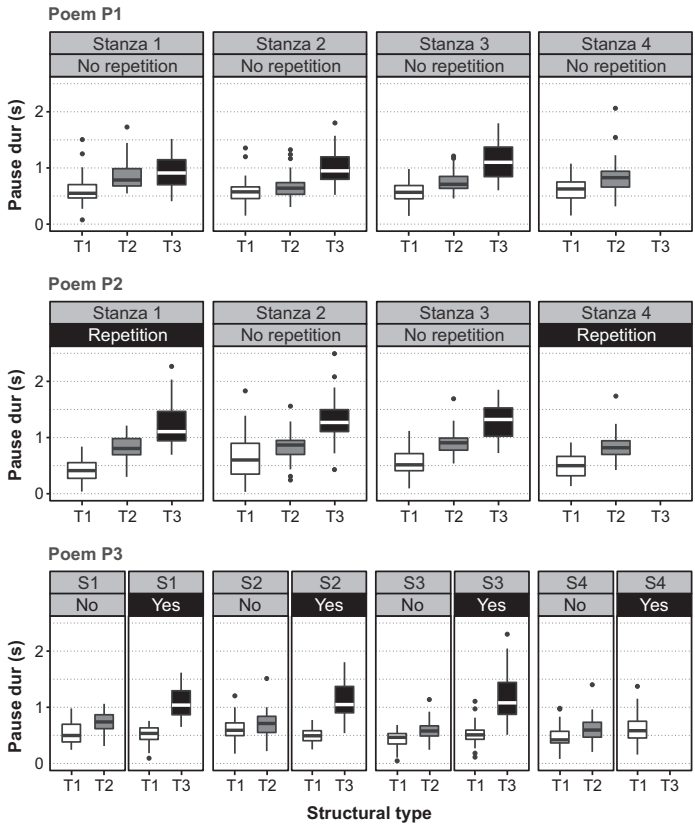


Figure 1 Duration of pauses (in seconds) as a function of STANZA and STRUCTURAL TYPE (T1: distich-initial line, T2: distich-final, stanza-non-final line, T3: stanza-final line). Black panels indicate repeated passages in P2 and P3, while grey panels indicate non-repeated passages.

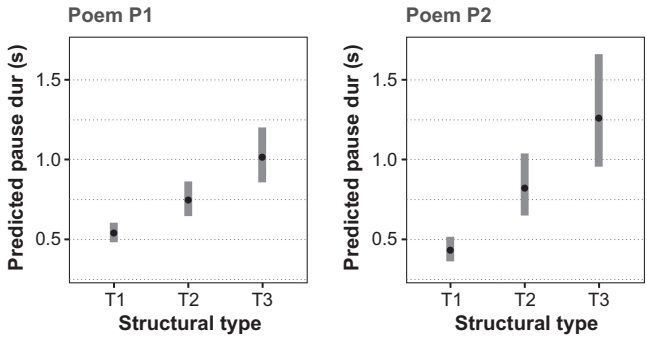


Figure 2 LME effect plot showing pause duration (back-transformed to seconds) as a function of STRUCTURAL TYPE (T1: distich-initial line, T2: distich-final, stanza-non-final line, T3: stanza-final line). Results are shown separately for poems P1 (left) and P2 (right).

Table 2 Tukey post-hoc comparisons of STRUCTURAL TYPE for pause duration, averaged over the levels of STANZA. Values represent duration ratios, with tests performed on the log scale. Results are shown for poems P1 and P2.

Model	Comparison	Estimate	Stan. Error	*Z*-ratio	*P*-value
P1	T1 / T2	0.723	0.058	−4.015	<.001
	T1 / T3	0.532	0.049	−6.862	<.001
	T2 / T3	0.736	0.076	−2.982	0.009
P2	T1 / T2	0.526	0.072	−4.693	<.001
	T1 / T3	0.343	0.053	−6.869	<.001
	T2 / T3	0.652	0.114	−2.454	0.042

In contrast to P1 and P2, a significant interaction between VERSE LINE and STANZA was found for P3 ($\chi^2(8) = 39.3$, $p < 0.001$). The corresponding effect plot is shown in Figure 3 (relevant post-hoc comparisons are reported in Tab. 3). In all stanzas (S1–S3), there was a robust difference between VL4 (= T3) and the other three verse lines, with statistically longer pauses at the ends of stanzas. However, unlike in the previous poems, VL2 (= T2) was significantly different from VL1 in all stanzas other than S2. However, the primary source of the interaction appeared to be the behaviour of VL3. In S1, pauses after VL3 were not significantly different from those after VL1, as expected, since both are of type T1. However, in S2, VL3 was associated with significantly shorter pauses than VL1, while in S3 and S4 the pattern was reversed.

A comparison of VL1 and VL2 (non-repeated text) with VL3 and VL4 (text repeated across stanzas) revealed no clear influence of repetition on pause duration. The fact that pauses in T3 contexts in the repeated passage were longer than T2 pauses in the non-repeated passage (= rows VL2/VL4 in Tab. 3) is consistent with the previous poems, and thus reflects structural effects rather than an effect of repetition.

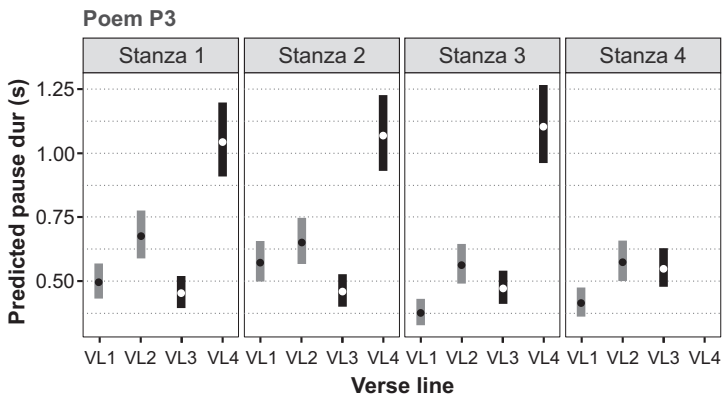


Figure 3 LME effect plot showing pause duration (back-transformed to seconds) as a function of STRUCTURAL TYPE (T1: distich-initial line, T2: distich-final, stanza-non-final line, T3: stanza-final line) and STANZA (S1–4), for poem P3.

Table 3 Tukey post-hoc comparisons of VERSE LINE within STANZA for pause duration for poem P3. Values represent duration ratios, with tests performed on the log scale.

Stanza	Comparison	Estimate	Stan. Error	*Z*-ratio	*P*-value
S1	VL1 / VL2	0.733	0.056	−4.038	< 0.001
	VL3 / VL4	0.433	0.033	−10.861	< 0.001
	VL1 / VL3	1.094	0.084	1.171	1.0
	VL2 / VL4	0.647	0.049	−5.652	< 0.001
S2	VL1 / VL2	0.879	0.067	−1.679	0.558
	VL3 / VL4	0.429	0.033	−10.996	< 0.001
	VL1 / VL3	1.247	0.096	2.865	0.025
	VL2 / VL4	0.609	0.046	−6.452	< 0.001
S3	VL1 / VL2	0.665	0.051	−5.292	< 0.001
	VL3 / VL4	0.425	0.032	−11.113	< 0.001
	VL1 / VL3	0.795	0.061	−2.980	0.017
	VL2 / VL4	0.508	0.039	−8.802	< 0.001
S4	VL1 / VL2	0.720	0.055	−4.271	< 0.001
	VL1 / VL3	0.754	0.058	−3.670	0.002

3.2 Articulation rate (syllables)

Figure 4 presents the syllabic AR of individual verse lines. In Poems P1 and P2, T2 lines were consistently delivered at a slower tempo than T1 lines. However, in P3, the opposite was the norm, except for the last stanza. T3 showed less consistent behaviour in P1 and P2, but in P3 it was notably stable and characterized by a fast AR. Regarding stanza-level trends, P1 did not show any descending tendency across stanzas, while P2 exhibited a gradual decline in AR from S1 to S3, followed by a reset in S4, which repeated the text of S1. In P3, the non-repeated lines (VL1 and VL2) showed a subtle downward trend in tempo across stanzas, while the repeated lines (VL3 and VL4) maintained a more consistent rate throughout.

In P1, there was no significant interaction between STRUCTURAL TYPE and STANZA ($\chi^2(6) = 12.6$, $p = 0.051$). However, including STRUCTURAL TYPE significantly improved the model fit ($\chi^2(2) = 18.9$, $p < 0.001$), as did including STANZA ($\chi^2(3) = 10.0$, $p = 0.019$). Figure 5 displays the predicted values from the linear model (STRUCTURAL TYPE on the left, STANZA on the right), while pairwise comparisons are summarized in Table 4. T2 lines were articulated significantly more slowly than T1 and T3, whereas there was no significant difference between T1 and T3. In contrast, only one significant pairwise comparison was found for STANZA (Tab. 3).

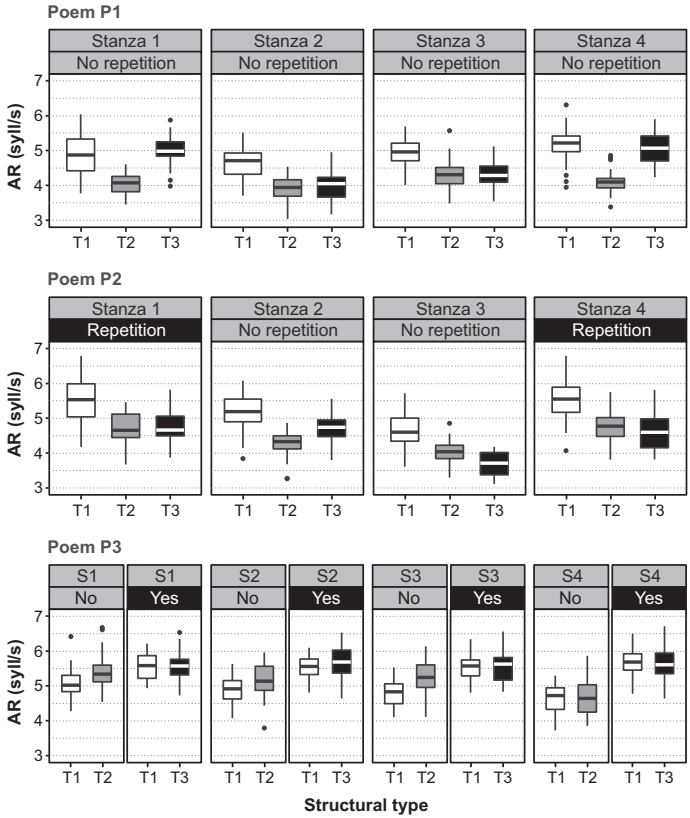


Figure 4 Articulation rate (in syllables per second) as a function of STANZA and STRUCTURAL TYPE (T1: distich-initial line, T2: distich-final, stanza-non-final line, T3: stanza-final line). Black panels indicate repeated passages in P2 and P3, while grey panels indicate non-repeated passages.

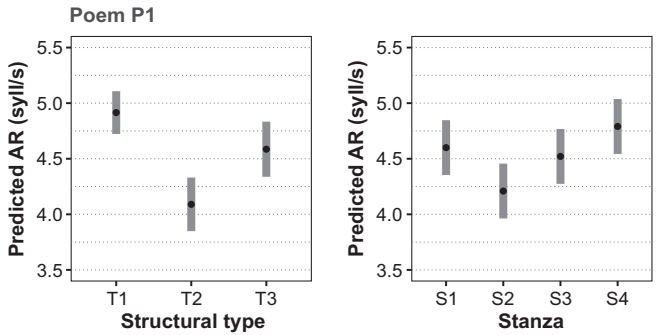


Figure 5 LME effect plot for AR (in syllables per second) in P1 (without interaction). Effect of structural type on the left (T1: distich-initial line, T2: distich-final line, T3: stanza-final line), effect of stanza on the right.

Table 4 Tukey post-hoc comparisons (difference of AR in syllables per second) for STRUCTURAL TYPE and STANZA for P1 (averaged over the levels of the other effect).

Main effect	Comparison	Estimate	Standard Error	*Z*-ratio	*P*-value
Structural type	T1 – T2	0.825	0.137	6.004	< 0.001
	T1 – T3	0.329	0.140	2.351	0.056
	T2 – T3	−0.496	0.161	−3.088	0.006
Stanza	S2 – S4	−0.581	0.159	−3.663	0.002

In P2, there was a significant interaction between STRUCTURAL TYPE and STANZA ($\chi^2(6) = 17.3$, $p = 0.008$). Focusing first on structural differences, T1 lines were articulated at a consistently faster rate than both T2 and T3 lines across all stanzas, with all comparisons reaching significance ($p < 0.001$). The contrast between T2 and T3 lines, however, was less reliable: it was not significant in S1 and S4 ($p > 0.05$), while it reached significance, but in opposite directions, in S2 (T2–T3 = −0.449, SE = 0.108, z-ratio = −4.168, $p < 0.001$) and S3 (T2–T3 = 0.335, SE = 0.108, z-ratio = 3.110, $p = 0.006$).

Comparisons of the same structures across stanzas (see Tab. 5) confirmed that AR generally decreased from S1 to S3. This decline was statistically significant at all steps, except for the S2–S3 transition for T2 lines and the S1–S2 transition for T3 lines. Importantly, for any VL type, there was no significant difference between S1 and S4 and, at the same time, S4 lines were significantly faster than the corresponding lines in S3 – highlighting a return to the initial tempo pattern.

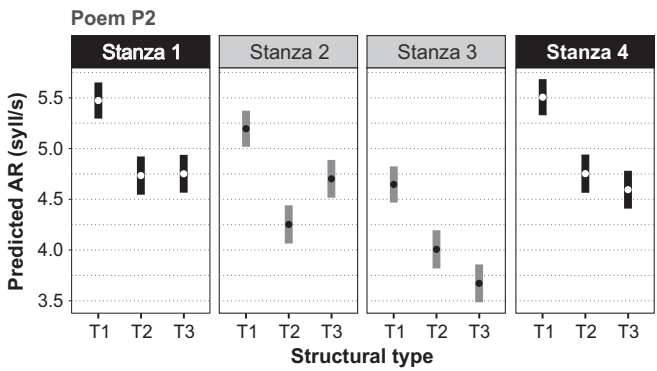


Figure 6 LME effect plot for AR (in syllables per second) in P2 as a function of STRUCTURAL TYPE (T1: distich-initial line, T2: distich-final, stanza-non-final line, T3: stanza-final line) and STANZA (1–4) in interaction. Darker shades indicate repeated passages (S4 identical to S1).

In P3, a significant interaction was found between VERSE LINE and STANZA ($\chi^2(9) = 108.2$, $p < 0.001$). Focusing on the effects of structural type, in the first distich, the difference between T1 and T2 was statistically significant in S1 to S3 but not in S4 (see Tab. 6). Namely, T2 verse lines were articulated significantly faster than T1, contrary to previous poems (where slower tempo occurred). In addition, regarding the second distich, there

Table 5 Tukey post-hoc comparisons (differences of AR in syllables per second) for STANZA within STRUCTURAL TYPE for P2.

Structural type	Comparison	Estimate	St. Error	*Z*-ratio	*P*-value
T1 (distich-initial line)	S1 – S2	0.278	0.075	3.730	0.001
	S2 – S3	0.547	0.075	7.334	< 0.001
	S3 – S4	−0.860	0.075	−11.534	< 0.001
	S1 – S4	0.035	0.075	−0.470	1.0
T2 (distich-final line)	S1 – S2	0.481	0.105	4.563	< 0.001
	S2 – S3	0.243	0.105	2.307	0.127
	S3 – S4	−0.746	0.105	−7.073	< 0.001
	S1 – S4	−0.021	0.105	−0.203	1.0
T3 (stanza-final line)	S1 – S2	0.050	0.105	0.474	1.0
	S2 – S3	1.027	0.105	9.746	< 0.001
	S3 – S4	−0.923	0.105	−8.754	< 0.001
	S1 – S4	0.155	0.105	1.466	0.856

was no significant difference between T1 and T3 in any stanza ($p > 0.05$). This means that the structural effect was limited to the non-repeated portion of the poem, and manifested in a reversed direction to that observed in P1 and P2.

Moreover, there was no significant effect of STANZA on the repeated VLs ($p > 0.05$), which were articulated at a similar tempo throughout the poem. In contrast, tempo tended to decrease in the non-repeated passages, namely, between S1 and S2 and between S3 and S4, but not between S2 and S3 (for pairwise comparisons, see Tab. 7). As a result, the difference between D1 and D2 gradually increased across stanzas (see the estimates for VL1 – VL3 and VL2 – VL4 in Tab. 6).

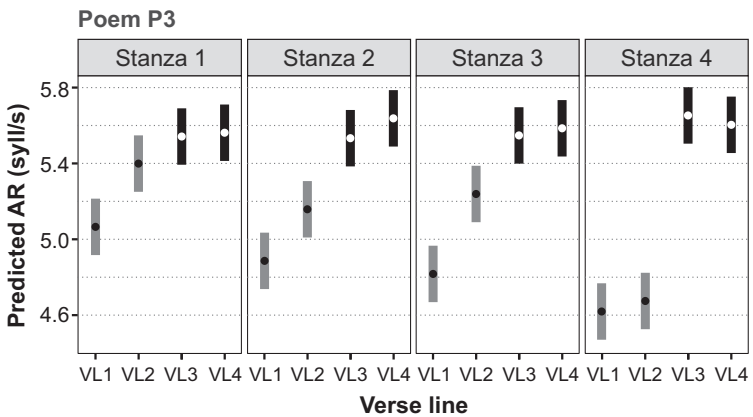


Figure 7 LME effect plot for AR (in syllables per second) in P3 as a function of VERSE LINE (lines 1–4, with VL2 corresponding to T2, VL4 to T3 in previous analyses) and STANZA (1–4) in interaction. Darker shades indicate repeated passages (identical distich appeared in VL3+VL4).

Table 6 Tukey post-hoc comparisons (differences of AR in syllables per second) for VERSE LINE within STANZA for P3 (comparisons VL3 – VL4 were not significant in any stanza).

Stanza	Comparison	Estimate	St. Error	*Z*-ratio	*P*-value
S1	VL1 – VL2	−0.334	0.067	−4.986	< 0.001
	VL1 – VL3	−0.476	0.067	−7.118	< 0.001
	VL2 – VL4	−0.163	0.067	−2.432	0.090
S2	VL1 – VL2	−0.272	0.067	−4.065	< 0.001
	VL1 – VL3	−0.647	0.067	−9.667	< 0.001
	VL2 – VL4	−0.4795	0.067	−7.166	< 0.001
S3	VL1 – VL2	−0.422	0.067	−6.299	< 0.001
	VL1 – VL3	−0.730	0.067	−10.914	< 0.001
	VL2 – VL4	−0.3465	0.067	−5.178	< 0.001
S4	VL1 – VL2	−0.055	0.067	−0.828	1.0
	VL1 – VL3	−1.034	0.067	−15.450	< 0.001
	VL2 – VL4	−0.929	0.067	−13.883	< 0.001

Table 7 Tukey post-hoc comparisons (differences of AR in syllables per second) for STANZA within VERSE LINE for the first distich in P3 (no such comparison was significant for the second distich, left out here).

Verse line	Comparison	Estimate	St. Error	*Z*-ratio	*P*-value
VL1	S1 – S2	0.178	0.067	2.663	0.046
	S2 – S3	0.072	0.067	1.071	1.0
	S3 – S4	0.198	0.067	2.961	0.018
VL2	S1 – S2	0.239	0.067	3.584	0.002
	S2 – S3	−0.077	0.067	−1.164	1.0
	S3 – S4	0.564	0.067	8.432	< 0.001

3.3 Articulation rate (phonemes)

Figure 8 presents the phonemic AR of individual verse lines. In P1, no consistent structural pattern can be identified. In P2, higher-level units (T2, T3) were associated with slower tempo than T1, with no obvious effect of repetition. In P3, a large effect of structural type is evident in the repeated lines (VL3 and VL4), and a smaller one in the non-repeated lines (VL1 and VL2), with higher levels being associated with slower tempo. The only obvious effect of repetition is that the repeated T1 lines were articulated at a higher phonemic AR than the non-repeated T1 lines.

In P1, there was no significant interaction between STRUCTURAL TYPE and STANZA ($\chi^2(6) = 12.0, p = 0.061$). None of the two fixed effects reached significance (STRUCTURAL TYPE: $\chi^2(2) = 2.85$, $p = 0.241$; STANZA: $\chi^2(3) = 3.9$, $p = 0.274$). Phonemic AR seems to

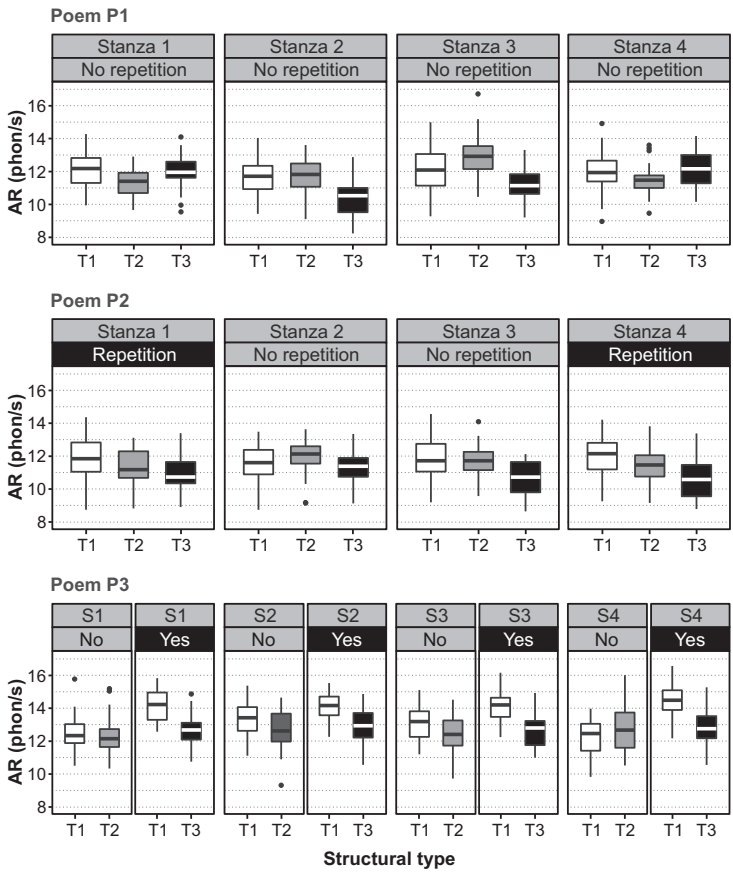


Figure 8 Articulation rate (in phonemes per second) as a function of stanza and STRUCTURAL TYPE (T1: distich-initial line, T2: distich-final, stanza-non-final line, T3: stanza-final line). Black panels indicate repeated passages in P2 and P3, while grey panels indicate non-repeated passages.

vary inconsistently across stanzas and types of verse lines in P1, leading to no significant effect.

Similarly, the interaction between STRUCTURAL TYPE and STANZA did not reach statistical significance in P2 ($\chi^2(6) = 11.8$, $p = 0.066$). However, this time the inclusion of STRUCTURAL TYPE significantly improved the model fit ($\chi^2(2) = 15.9$, $p < 0.001$), although this was not the case for the inclusion of STANZA ($\chi^2(3) = 0.3$, $p = 0.964$). Specifically, the final lines in a stanza were articulated at a significantly slower phonemic rate than the distich-initial lines (T1–T3 = 0.976, SE = 0.197, z-ratio = 4.940, $p < 0.001$) and the distich-final VL2 (T2–T3 = 0.716, SE = 0.213, z-ratio = 3.358, $p = 0.002$). There was no significant difference between T1 and T2 (T1–T2 = 0.259, SE = 0.190, z-ratio = 1.364, $p = 0.518$). The effect plot is shown in Figure 9. Phonemic AR was comparable across stanzas, yielding no effect of repetition in S4.

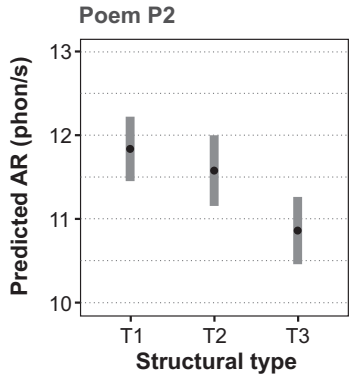


Figure 9 LME effect plot for AR (in phonemes per second) in P2 as a function of STRUCTURAL TYPE (T1: distich-initial line, T2: distich-final, stanza-non-final line, T3: stanza-final line).

In P3, there was a significant interaction between VERSE LINE and STANZA ($\chi^2(9) = 60.6$, $p < 0.001$). Figure 10 displays the interaction plot. Focusing on effects of structural type, in the first two lines, the difference between T1 and T2 was statistically significant in all stanzas but S1 (see Tab. 8). Namely, T2 verse lines were articulated significantly slower than T1 in S2 and S3 but faster than T1 in S4. Regarding the second distich, there was a consistent difference between T1 and T3, with the stanza-final lines being articulated significantly slower than the VL3. This suggests that the structural effect was relevant chiefly for the repeated parts of the poem, and appeared less consistently for the non-repeated parts. Importantly, the difference between repeated and non-repeated passages emerged only in the initial line of the distichs: VL3 was articulated significantly faster than VL1, while no comparable difference was observed in the final lines (see Tab. 8).

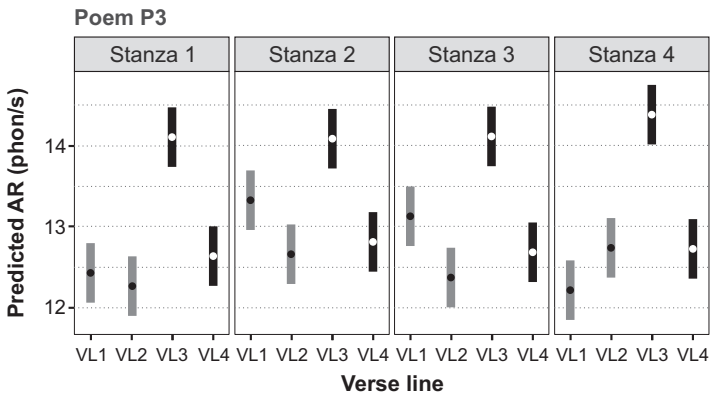


Figure 10 LME effect plot for AR (in phonemes per second) in P3 as a function of VERSE LINE (lines 1–4, with VL2 corresponding to T2, VL4 to T3 in previous analyses) and stanza (1–4) in interaction. Darker shades indicate the repeated part of the poem (identical distich in VL3+VL4).

Table 8 Tukey post-hoc comparisons (differences of AR in phonemes per second) for VERSE LINE within STANZA for P3.

Stanza	Comparison	Estimate	Stan. Error	*Z*-ratio	*P*-value
S1	VL1 – VL2	0.163	0.164	0.991	1.0
	VL3 – VL4	1.466	0.164	8.919	< 0.001
	VL1 – VL3	−1.673	0.164	−10.180	< 0.001
	VL2 – VL4	−0.370	0.164	−2.251	0.146
S2	VL1 – VL2	0.665	0.164	4.048	< 0.001
	VL3 – VL4	1.272	0.164	7.737	< 0.001
	VL1 – VL3	−0.758	0.164	−4.612	< 0.001
	VL2 – VL4	−0.152	0.164	−0.923	1.0
S3	VL1 – VL2	0.755	0.164	4.594	< 0.001
	VL3 – VL4	1.427	0.164	8.683	< 0.001
	VL1 – VL3	−0.983	0.164	−5.983	< 0.001
	VL2 – VL4	−0.311	0.164	−1.894	0.349
S4	VL1 – VL2	−0.571	0.164	−3.474	0.003
	VL3 – VL4	1.654	0.164	10.062	< 0.001
	VL1 – VL3	−2.212	0.164	−13.458	< 0.001
	VL2 – VL4	0.013	0.164	0.078	1.0

There was no significant effect of STANZA on the repeated verse lines VL3 and VL4, but also in the non-repeated VL2 ($p > 0.05$). Phonemic AR differed significantly across stanzas only for VL1: tempo increased between S1 and S2 (S1–S2 = −0.896, SE = 0.164, z-ratio = −5.450, $p < 0.001$), did not differ between S2 and S3 ($p > 0.05$), decreased between S3 and S4 (S3–S4 = 0.960, SE = 0.164, z-ratio = 5.842, $p < 0.001$).

4. Discussion

This study investigated how information structure – specifically the distinction between given and new information – and poetic structure interact to shape the temporal aspects of poetic delivery. Three structurally similar poems were selected to provide a controlled context for examining the effects of textual repetition and stanza structure. We examined how performers modulate pause duration and articulation rate (AR), both syllabic and phonemic, in response to these factors. Repeated lines, representing given information, were typically articulated at a faster rate and with more stable prosodic timing than non-repeated lines, which introduce new information. The findings demonstrate that performers consistently adjust temporal features to reflect informational status, even within the rhythmic and structural constraints of poetry.

Such adaptations not only serve to highlight informational prominence and guide listener attention but also contribute to the perceptual marking of the poem's textual architecture.

Poem 1: Baseline without repetition

P1, containing only non-repeated (new) lines, served as a control condition. As anticipated, no systematic variation in AR was observed across stanzas, suggesting the absence of repetition precludes consistent tempo adjustments. Variability in tempo across verse lines appeared to stem from local syntactic or lexical complexity rather than from structural positioning. However, pause duration robustly marked structural divisions: pauses were longest at stanza boundaries (T3), intermediate at mid-stanza breaks between distichs (T2), and shortest between the two lines within each distich (T1). Although tempo variation was a less reliable cue to this structure, T1 lines were also generally articulated at the fastest rates, while T2 and T3 lines were mostly associated with a decrease in tempo, which aligns with prior findings on prosodic boundary signalling (the so-called *final lengthening/deceleration*).

Poem 2: Full-stanza repetition

P2 introduced a full-stanza repetition: the final stanza was identical to the first. The expected prosodic cues for structural boundaries were again evident, with pause durations reliably distinguishing between T1, T2, and T3 positions. In contrast to P1, tempo patterns were this time more structured: syllabic AR tended to be highest in T1 lines and lowest in T3 lines, with T2 lines occupying an intermediate position. The final line is each stanza was also articulated significantly slower than T1 and T2 lines in terms of the phonemic AR.

A notable finding was the global declination of syllabic AR across the first three stanzas, followed by a tempo reset in the repeated stanza (S4), returning to the tempo of the initial stanza. This pattern thus mirrors prosodic phrasing in speech, where declining melody or tempo can reset at phrase boundaries (cf. Volín et al., 2024). Contrary to our expectation that repeated content would be delivered faster than its original occurrence, the repeated stanza was not faster than S1 – but it was significantly faster than both S2 and S3. This suggests a deliberate strategy: performers slow down progressively to prepare for signalling repetition via a tempo reset, rather than by accelerating the repeated lines themselves.

Poem 3: Distich-level repetition

P3 introduced a distinct repetition pattern: each stanza's second distich repeated across all stanzas, while the first distich remained unique. As predicted, the repeated lines (VL3 and VL4) were articulated at a consistently higher syllabic AR than the non-repeated lines (VL1 and VL2). However, although the contrast became more pronounced across stanzas, in line with our predictions, this was due to a tempo decline in the non-repeated section, not an increase in the repeated one (the repeated passage maintained a uniform tempo in all stanzas). This in fact replicates the pattern from P2, suggesting that performers use temporal declination in new content to create perceptual contrast against stable, and thus faster delivery of repeated material.

An unexpected result emerged in the first stanza: the repeated distich (D2) was delivered more quickly than the new distich (D1), even though it was the first appearance of both. A plausible explanation is that the speakers were already familiar with the text from prior rehearsal, during which they read each new D1 once, but D2 four times, effectively reclassifying it as given information in the first stanza despite its initial mention during recording. No tempo acceleration occurred across later iterations of D2, indicating that familiarity had plateaued.

General observations on temporal structuring

Across all poems, the typology of verse lines (T1, T2, T3) influenced temporal delivery. T1 lines (within distichs) were generally the fastest, while T2 (distich-final) and T3 (stanza-final) lines were slower and accompanied by longer pauses. Importantly, this pattern persisted even in P3, where all lines had equal syllable counts (11 syllables), unlike P1 and P2, where the T2 and T3 lines were shorter (10 syllables). This counters the possibility that tempo differences arise solely from line length (cf. Quené, 2005). Instead, it supports a structural explanation: unit-finality is associated with slower delivery and longer pauses, likely due, in part, to both syntactic closure and prosodic boundary marking.

Similarly, the stepwise decrease in AR across stanzas may serve as a general acoustic cue for stanza positioning, allowing listeners to infer where in the poem the speaker currently is based on tempo. However, since this pattern was absent in the baseline poem (P1), it is more plausibly interpreted as a deliberate strategy to signal upcoming repetition rather than stanza position alone.

Finally, it is important to consider how the choice of measuring AR primarily in syllables rather than phonemes influenced our results. While syllabic and phonemic AR were strongly correlated in all poems (P1: $r = 0.75$ [0.71, 0.78]; P2: $r = 0.74$ [0.70, 0.78]; P3: $r = 0.78$ [0.75, 0.82]; $p < 0.001$), they also exhibited distinct patterns, suggesting that each captures different aspects of speech tempo. We chose to calculate AR based on syllables because syllable counts were controlled across verse lines, unlike phoneme counts, which varied considerably and randomly due to differences in phonotactic complexity. In Czech, the presence of frequent consonant clusters (Šturm & Bičan, 2021) can lead to fluctuations in phonemic AR unrelated to structural or informational factors. Syllabic AR, by contrast, offers a more consistent approximation of perceived tempo when syllable length is held constant.

This distinction helps explain why phonemic AR did not vary systematically in P1, despite structural changes. In P2, phonemic AR was lower at stanza endings – consistent with syllabic AR – but remained stable across stanzas, showing no sensitivity to repetition. The most revealing discrepancies emerged in P3. In the non-repeated distichs, T2 lines were faster than T1 lines in syllabic AR but slower in phonemic AR (only the latter aligns with our expectations). Similarly, in the repeated distichs, syllabic AR showed virtually no difference between T1 and T3, whereas phonemic AR revealed a marked slowing in T3.

These results suggest that the most perceptually accurate measure of tempo might involve a composite metric combining syllabic and phonemic AR (cf. Pfitzinger, 1998). While the former reflects listener-perceived rhythm, the latter captures articulatory density. The two are not interchangeable, and their independent behaviour points to a more comprehensive model of perceived tempo that integrates both measures.

Acknowledgements

This work was funded by the Czech Science Foundation (GAČR), project 24-10905S: *Prosodic expression of utterance information structure in Czech*. The author would like to thank Leona Straková and Barbora Vavřičková for help with the segmentation of part of the material that was used in their BA theses, focusing on the variability of nasals in poem P2 (Straková, 2023) and vocalic intervals in P3 (Vavřičková, 2023). Sincere thanks should also be given to Jan Volín for his continuous support.

REFERENCES

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

Campione, E., & Véronis, J. (2002). A large-scale multilingual study of silent pause duration. In: *Proceedings of Speech Prosody 2002* (pp. 199–202).

Crystal, T. H., & House, A. S. (1990). Articulation rate and the duration of syllables and stress groups in connected speech. *Journal of the Acoustical Society of America*, *88*, 101–112.

Eefting, W. (1991). The effect of 'information value' and 'accentuation' on the duration of Dutch words, syllables and segments. *Journal of the Acoustical Society of America*, *89*, 412–424.

Firbas, J. (1979). A functional view of 'ordo naturalis'. *Brno studies in English*, *13*(1), 29–59.

Fowler, C. A., & Housum, J. (1987). Talkers' signaling of 'new' and 'old' words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language*, *26*, 489–504.

Chafe, W. (1974). Language and consciousness. *Language*, *50*, 111–133.

Lenth, R. (2022). *emmeans: Estimated marginal means, aka least-squares means* [R package version 1.8.2]. Available at <https://CRAN.R-project.org/package=emmeans>.

Lieberman, P. (1963). Some effects of semantic and grammatical context on the productions and perception of speech. *Language and Speech*, *6*, 172–187.

Nooteboom, S. G., & Eefting, W. (1994). Evidence for the adaptive nature of speech on the phrase level and below. *Phonetica*, *51*, 92–98.

Paschen, L., Fuchs S., & Seifart, F. (2022). Final lengthening and vowel length in 25 languages. *Journal of Phonetics*, *94*, Art. 101179.

Pfitzinger, H. R. (1998). Local speech rate as a combination of syllable and phone rate. *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP 1998)*, Art. 0523.

Quené, H. (2005). Modeling of between-speaker and within-speaker variation in spontaneous speech tempo. In: *Proceedings of Interspeech 2005* (pp. 2457–2460).

R Core Team (2022). *R: A language and environment for statistical computing* [Computer program, version 4.2.1]. R Foundation for Statistical Computing, Vienna, Austria. Available at <https://www.R-project.org>.

Straková, L. (2023). *Variabilita vybraných akustických vlastností nazál při recitaci básně* [Unpublished B.A. thesis]. Charles University, Faculty of Arts.

Svoboda, A. (1981). *Diatheme: A study in thematic elements, their contextual ties, thematic progressions and scene progressions based on a text from Aelfric*. Univerzita J. E. Purkyně.

Šturm, P., & Bičan, A. (2021). *Slabika a její hranice v češtině*. Karolinum.

Šturm, P., & Volín, J. (2023). Occurrence and duration of pauses in relation to speech tempo and structural organization in two speech genres. *Languages*, *8*(1), Art. 23.

Trouvain, J., Koreman, J., Erriquez, A., & Braun, B. (2001). Articulation rate measures and their relation to phone classification in spontaneous and read German speech. In *Proceedings of the Workshop on Adaptation Methods for Speech Recognition* (pp. 155–158).

Vavřičková, B. (2023). *Variabilita trvání vokalických intervalů při recitaci básně* [Unpublished B.A. thesis]. Charles University, Faculty of Arts.

Veroňková-Janíková, J. (2004). Dependence of individual speaking rate on speech task. *AUC Philologica, 2004*(1), 79–95.
Volín, J., Šturm, P., Skarnitzl, R., & Bořil, T. (2024). *Prosodic phrase in spoken Czech*. Karolinum.
Volín, J. (2022). Variation in speech tempo and its relationship to prosodic boundary occurrence in two speech genres. *AUC Philologica*, *2022*(1), 65–81.
Werner, R., Trouvain, J., & Möbius, B. (2022). Optionality and variability of speech pauses in read speech across languages and rates. In *Proceedings of Speech Prosody 2022* (pp. 312–316).
Zellner, B. (1994). Pauses and the temporal structure of speech. In E. Keller (ed.), *Fundamentals of speech synthesis and speech recognition* (pp. 41–62). John Wiley & Sons.

APPENDIX

Poem P1
Viktor Dyk: 'Sentimentální balada'

I on ví: V království předků to kdesi,
uprostřed lesů, hor stojí ten hrad.
Není tam příšery, která tak děsí…
A kdo tam pronikne, umí se smát!

Smáti se do oblak, která tam čistá!
Smáti se v komnatách, kde bol vždy ztich!
Smáti se všemu, co vzrůsti se chystá!
Dětinský, veselý, volný ten smích!

To on ví. V paláci trvá však dále
uprostřed hladkých svých dvořanů řad.
Království spravuje ku Boží chvále.
Moudrý je, slavný je – nezná se smát!

A kdyby odešel, v nový cíl věře,
ví, hrad ten zaklel by démon mu pryč…
Devíti zámky by zavřel mu dvéře,
do řeky hodil by železný klíč!

Poem P2
Viktor Dyk: 'Na melodii neznámé písně'

Ty tóny duši rozrývají maně
a pohádka to promrskaná dost.
– Já o jedné jen sníval karavaně
a na poušti se bělá její kost –!

Je vzdálena a cesty neznámy mi.
Jen písek zříš, když hledíš do dáli!
A slunce líbá rety žíznivými
ty, kteří kostrou nyní zůstali…

A při rozmarném paprsků těch tanci
teď vzpomínám, co žilo v kostrách těch.
Jich táhlou vzpomínám já na romanci,
jich hořký smích a galantní jich vzdech.

Ty tóny duši rozrývají maně
a pohádka to promrskaná dost.
– Já o jedné jen sníval karavaně
a na poušti se bělá její kost –!

Poem P3:
František Gellner: 'XXXI.'

V kavárně u stolku lecco se řekne,
srdce se zachvěje, srdce se lekne.
Trochu se vraždilo, trochu se kradlo,
pereme, pereme špinavé prádlo.

Otec tvůj poslední prodal již krávu,
matku bůh povolal ve svoji slávu.
Trochu se vraždilo, trochu se kradlo,
pereme, pereme špinavé prádlo.

Slova jsou slova a mladost je mladost,
genitálie si přejí svou radost.
Trochu se vraždilo, trochu se kradlo,
pereme, pereme špinavé prádlo.

Ve zraku holek plá nemilá tklivost,
hostinských zmáhá se netrpělivost.
Trochu se vraždilo, trochu se kradlo.
Pereme, pereme špinavé prádlo.

RESUMÉ

Studie se zabývá otázkou, jak prozodické prostředky odrážejí informační strukturu a básnickou organizaci při recitování poezie. Zaměřuje se na vliv opakování a hierarchické struktury verše na artikulační tempo a trvání pauz. Výzkumu se zúčastnilo 32 rodilých mluvčích češtiny, kteří přednesli několik básní, z nichž byly vybrány tři formálně obdobné básně lišící se mírou a rozmístěním opakovaných veršů. První báseň neobsahovala žádné opakování (kontrolní vzorek), v druhé básni se opakovala celá sloka a třetí báseň obsahovala opakované dvojverší v rámci každé sloky. Analýza ukázala, že opakované verše (považované za danou informaci) byly produkovány rychleji a s menší variabilitou než verše nové (s novou informací). Ve strukturách s opakováním se navíc objevoval postupný pokles tempa s následným obnovením původních hodnot, což naznačuje záměrné využívání modulace tempa ke zdůraznění

textového opakování. Trvání pauz spolehlivě vyznačovalo hranice vyšších strukturních celků, přičemž nejdelší pauzy byly zaznamenány na rozhraní slok. Rozdíly mezi slabičným a fonémickým artikulačním tempem dále poukazují na vliv fonotaktické variability češtiny. Výsledky celkově potvrzují, že mluvčí aktivně využívají temporální aspekty k vyjadřování informačních i strukturních vztahů v textu, čímž podporují srozumitelnost a vnímání básnické formy.

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

2025 ACTA UNIVERSITATIS CAROLINAE PAG. 85–95
PHILOLOGICA 3

PROSODIC PROMINENCE OF MODAL VERBS IN NARRATIVES

JAN VOLÍN

ABSTRACT
The study deals with accentuation (presence or absence of realized lexical stress) of modal verbs in continuous spoken texts. The material was taken from audiobooks where Czech professional actors (8 male + 8 female) read out various narratives produced by renowned authors. A corpus of over 17,000 words was used. The recordings were annotated with special attention to presence or absence of materialized stress (i.e., accent). The chief purpose of the study was to provide descriptive data on modal verbs in connected speech, namely in the genre of narrative monologue. The results showed that modal verbs were deaccented in more than a quarter of their occurrences, but that monosyllabic and polysyllabic verbs behaved differently from each other. The infinitives associated with modal were also inspected. They were much less often unaccented and, importantly, the influence of monosyllabicity followed a different pattern than in modals. Additionally, information on mutual position of modals and associated infinitives is provided and an observation of negative forms of modals is made. The data can be further used in follow-up research to find out how structural and communicative requirements interact to produce the actual prosodic forms, or how modals in narratives differ from those in other communicative genres.

Keywords: accenting; associated infinitive; modal verbs; modality; negative form; prominence; prosodic backgrounding

1. Introduction

Like in many other languages, the lexical *stress potential* in Czech is preferably materialized in the actual use on autosemantic words (nouns, adjectives, full verbs, etc.), whereas synsemantic words (pronouns, conjunctions, auxiliary verbs, etc.) are candidates to remain unaccented or prosodically backgrounded (e.g., Palková, 1994: 282; Volín et al., 2024: 34–36). If not specified otherwise in this study, the term *stress* will refer to a prominence *potential* of the first syllable in Czech lexical items, whereas the term *accent* will refer to actually *materialized* prominence in spoken utterances (see also Volín & Skarnitzl, 2020).

Contrary to the unrefined rule mentioned in the preceding paragraph, a recent study revealed that a considerable number of autosemantic verbs (also termed full verbs) are

https://doi.org/10.14712/24646830.2025.22

not accented in Czech narratives (Volín & Hanžlová, 2024, but *cf.* also Franz et al., 2022 on German). These cases (over 10% of all autosemantic verbs), however, were plausibly explained if information structure was taken into consideration. The concept of *givenness* seemed to be especially useful in justifying the 'deaccenting'. When the semantic content of the verb was fully or partially given by the preceding co-text or factual context, the stress potential of the verb was not exploited by the speaker and the verb remained unaccented.

Unmarked unaccentedness, on the other hand, was typical of auxiliary verbs that served as mere grammatical markers of Czech past or future tense, or of verbal copulas (Volín & Hanžlová, 2024). These facts inspired a question concerning another principal group of verbs: the *modal verbs* or *modals* for short. They are often classified as auxiliary, yet with some reservations since they typically signify important concepts of *ability*, *obligation*, *possibility*/*probability* or *permission* (e.g., Grepl & Karlík, 1986; Klinge, 1993; Hoye, 2005). These subjective evaluations of the speakers' views on reality seem to be less predictable and more informative than semantic contents of typical auxiliaries. The prosodic prominence of modal verbs is, therefore, a chief concern in this study. We would like to find out, whether the extent of (un)accentedness in modal verbs is more like that of auxiliary verbs or that of full verbs, or whether it is similar to neither of the two types.

There is no consensus in the exact delimitation of the set of modal verbs even within one language, let alone across languages, even if scholars generally agree on what modality is (see, e.g., Svoboda, 1967; Benešová, 1973; van der Auwera & Plungian, 1998; Palmer, 2001; Hengeveld, 2004; Nauze, 2008). It follows that certain modal verbs are recognized as such even by authors who otherwise disagree with each other in various conceptual standpoints. Disagreements usually involve classification schemes. Typically, there is a small set of 'core' modal verbs like *must* (*muset* in Czech), may (*smět* in Cz.), can (*moct* or *umět* in Cz.), and an extended set in which verbs of similar meanings but different grammatical or semantic properties are included. Our present study has no ambition to contribute to the debate on what exactly modal verbs are. We use the guidance of Karlík and Šimík (2017) and, for interested parties, we present all the verbs included in our analyses in the Appendix at the end of this study. Rather, our concern is accentuation of the modal verbs and a few related questions, like that of infinitives associated with modals in an utterance (see below).

As our study is exploratory and not confirmatory, we stipulate no hypotheses. Instead, seven research questions are asked. These are listed below and the presentation of the results in this paper will refer to them as they are numbered here.

Research Questions

RQ1: How often is the modal verb in continuous narratives accented/unaccented?
RQ2: What is the ratio of monosyllabic modal verbs in the accented and unaccented set?
RQ3: What is the ratio of accented and unaccented infinitives associated with the modals?
RQ4: Does the monosyllabicity in associated infinitives display similar pattern as in modals?
RQ5: How often is the modal verb followed immediately by the associated infinitive?
RQ6: What is the accentuation pattern if the modal verb immediately precedes the infinitive?
RQ7: Do negative forms of modals display similar patterns as the positive forms?

Monosyllabicity in RQ2 and RQ4 is of interest because of the Stress Clash Rule, discussed sometimes as the Stress-Class Resolution or Rhythm Rule (e.g., Geigerich, 1992: 195; Hualde, 2010; Féry, 2017: 215; Lunden, 2019: 77; for Czech see Volín & Skarnitzl, 2018: 66). This rule operates in Czech and constrains occurrences of two accented syllables neighbouring each other (with some exceptions). It follows that monosyllabic words in phrase-internal position need specific treatment to avoid stress clash.

In sum, the purpose of this exploratory 'mapping' is to provide a clearer picture of the prosodic situation regarding modal verbs should any specialized future research necessitate this elementary information (see also Section 4 – Discussion).

2. Method

2.1 Material

Our material represents the genre of narratives read out to audiences. Even though the target listeners were not facing the speaker at the time of the reading, the recordings represent speech with clear communicative objectives. They were produced in studios manufacturing audiobooks for commercial purposes. The speakers were theatre or film actors by profession with established reputation for their skills. Likewise, the texts of the narratives were written by renowned authors. Given the purpose of the recordings, great care can be presumed in the production of the narratives (for instance, a presence of a director, follow-up checking, perhaps even pre-production consultations). We, therefore, believe that our material represents ecologically valid speech performances of the given genre.

As to the extent of the sample, there were 16 speakers (8 male + 8 female), and we required a stretch of continuous spoken text of at least 1,000 words per speaker. That amounts to approximately 140 utterances or 240 prosodic phrases per speaker. The speakers were of various ages ranging from about 30 to 65 years of age. Likewise, the time of recording spans about fifty years, from the 1970s. Therefore, we can generally speak of adult professional speakers of current Czech.

2.2 Sample annotation

Annotation of the sound recordings was carried out in Praat (Boersma & Weenink, 2022). It was all done in a blind fashion as a general part of corpus construction project at the Institute of Phonetics in Prague. It follows that the annotators did not know any hypotheses concerning the later exploitation of the corpus. Careful manual annotations (preceded by some semiautomatic procedures) were done at the level of phones, accent-groups, prosodic phrases and utterances. Within accent-groups (i.e., actually materialized stress-groups) each syllable was marked as accented or unaccented. All annotations were done by trained phoneticians with senior experts always checking the whole process. It should be noted that in the case of accent status identification, two senior experts reached mutual agreement of 97.5%. The rest (2.5% of the cases) was resolved through a debate over the repeated listening to audio recordings. Modal verbs were identified by the author of this study with the guidance of Karlík & Šimík, 2017.

3. Results

The sample provided 236 modal verbs. Table 3.1 provides their overview: central meanings with the number of occurrences in the sample. (A complete list of the Czech verbs identified in the spoken text here can be found in the Appendix at the end of this article.) The verb *can* is divided into two sets according to the meaning because apart from the central meaning of probability or permission, the Czech language has the verb *umět* which is very often translated by the English *can*. This division, however, was performed just for this introductory overview. It is not utilized in the analyses of prominence since it would open a complex area of research that would reach beyond the scope of this paper.

Table 3.1 English equivalents of modal verbs found in the sample with the numbers of their occurrences.

Verb	*n*
must	82
can (permission)	54
can (ability)	46
want	43
may	6
have sth. done	5

The answer to RQ1 (i.e., *how often is the modal verb accented*) is displayed in Figure 1. Almost a third (29.2%) of the modal verbs found in our sample were realized without an accent. That is clearly different from what previous research found in autosemantic verbs on the one hand, and auxiliary verbs on the other hand.

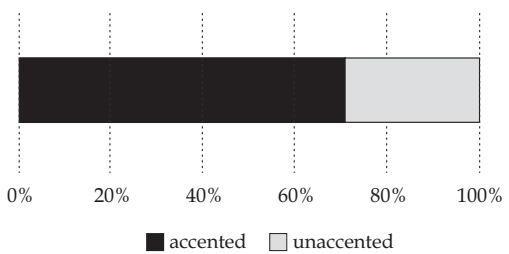


Figure 1 Visualization of the ratio of accented (black colour) and unaccented (grey colour) modal verbs in the sample of narratives.

To see how rhythmic and communicative requirements interact, it is important to examine the ratio of monosyllabic verbs in the accented and unaccented sets. This is because Stress-Clash Rule in Czech constrains accentuation of monosyllables in phrase-internal positions (Volín & Skarnitzl, 2018: 66). Figure 2 displays the ratios, which in effect provides the answer to RQ2 (i.e., *occurrence of monosyllabic modals in the accented and unaccented set*).

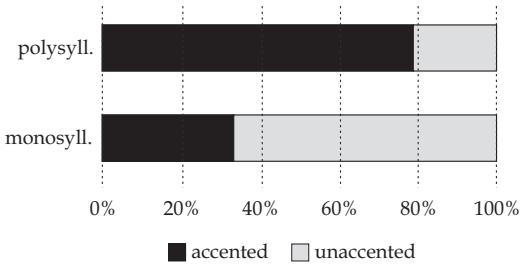


Figure 2 The ratios of accented (black colour) and unaccented (grey colour) modal verbs in the set of polysyllabic verbs (top bar) and monosyllables (bottom bar).

The difference in the ratios was established as statistically significant: χ^2 (1) = 34.6; $p < 0.001$, hence more or less holding in other samples of narratives outside the current one. Apparently, the monosyllabic status strongly increases the probability of 'deaccenting'. The fact that two thirds (66.7%) of monosyllables are unaccented, whereas only 21.1% of polysyllables were produced without an accent suggests that the semantic importance and informational status of the modal verbs are not that powerful – rhythmic considerations, namely the SCR (Stress-Clash Rule), exert their influence, too.

One of the defining features of modal verbs is their obligatory link to infinitives of other verbs whose semantic scope is affected by the modals. The result in the previous paragraph leads to a related question – that of *accentedness in associated infinitives* (RQ3). Although we found 236 modal verbs, there were only 227 associated infinitives. In nine instances, the infinitive was elided as it was obvious from the co-text and did not have to be repeated. Arguably, such elisions could be interpreted as ultimate backgrounding and, therefore, added to the count of unaccented (hence backgrounded) infinitives. However, to present descriptive data in clearer fashion, the nine infinitive elisions are held apart and excluded from the following graphs.

Figure 3 visualizes the ratio of accented to unaccented infinitives (associated with the modal verbs). There were 35 (15.4%) unaccented and 192 (84.6%) accented infinitives. Given that modal verbs merely adjust their infinitives, this result is not surprising. Also, it corresponds with what Volín and Hanžlová (2024) found out about autosemantic verbs – majority of them were accented with a very similar ratio.

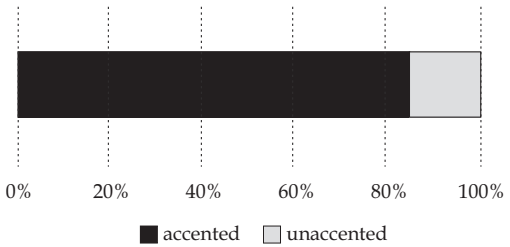


Figure 3 The ratio of accented (in black) and unaccented (in grey) infinitives associated modal verbs.

What needs to be asked now, is the ratio of unaccented to accented monosyllables and polysyllables in the infinitive set, i.e., a question parallel to RQ2 (and Figure 2 which displays the answer to it). In the Introduction, the question of unaccented monosyllables and polysyllables in infinitives is labelled RQ4. Figure 4 shows that monosyllabicity plays a role: only 5% of polysyllabic infinitives were produced without an accent. The difference in accentedness between monosyllabic and polysyllabic infinitives is statistically significant: χ^2 (1) = 58.2; $p < 0.001$.

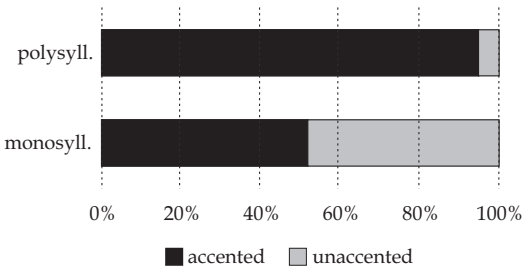


Figure 4 The ratios of accented (in black) and unaccented (in grey) infinitives in polysyllabic (top bar) and monosyllabic (bottom bar) forms.

Figures 2 and 4 are worth comparing: in the case of monosyllabic modals, the ratio of unaccented to accented items is 2 : 1, whereas in monosyllabic infinitives it is roughly 1 : 1. In polysyllabic modals, the ratio of unaccented cases to accented items is about 1 : 4, whereas in infinitives it is merely 1 : 19. Entering the actual counts (the outcomes of RQ2 and RQ4) into a statistical significance test yields the following result: χ^2 (7) = 25.6; $p < 0.001$. Inspection of the individual test criterion components shows that the main contributor to the significance is the behaviour of unaccented polysyllables in modals and infinitives.

Research questions 5 and 6 ask about mutual positions of modal verbs and infinitives and their stress pattern. A pictorial answer to RQ5 is offered in Figure 5. There were 149 (66.2%) infinitives immediately following their modal verb, and 76 (33.8%) infinitives with one or more words intervening. In other words, about two thirds of modal verbs are immediately followed by their associated infinitive in Czech narratives.

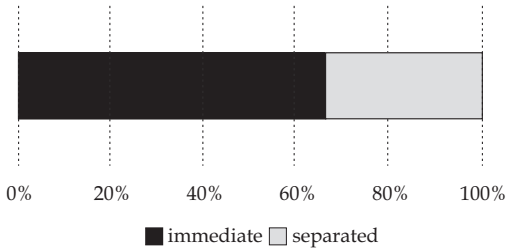


Figure 5 The ratio of infinitives immediately after modals (in black) to infinitives separated from their modals with one or more words (in grey).

The nine cases of infinitive elisions are not included and neither are two cases of infinitives preceding the modal. However, these two exceptional pre-positioned cases should

perhaps be presented here: exceptions often point to interesting aspects of the problem. Here, the relevance is guaranteed by a clear link to prominence/backgrounding issues and information structure considerations. The first wording is as follows:

(1) *Spát se nám nechtělo ani trochu.* *To sleep we didn't want a least bit.*

The infinitive here is in the initial position because it represents the theme of the utterance. It is unambiguously contextually bound (Daneš, 1979). The co-text before this utterance describes in some detail an evening situation and location that is clearly set for sleeping. The actors (a group of children) are getting ready to sleep. Moreover, the verb *sleep* itself is actually uttered 12 words beforehand. The phrase '*didn't want a least bit*' is apparently the rheme of the utterance and its position after the theme is typical of Czech information structure.

The second case is somehow different:

(2) *…, což jsem ovšem tušit nemohl* *…, which I actually suspect could not*

The infinitive *tušit* (*to suspect*) is not implicated by the co-text and does not represent the theme (the theme comprises the subject *I* referred to by the morphemic *-em* of *jsem*, and a fact referred to by the relative pronoun *což*). However, the post-positioning of the modal verb together with the nuclear accent creates extra strong emphasis on the expressed ability, or in this case, inability of the narrator. That gives the utterance more affective power.

Be that as it may, we still have 225 modal verbs with their associated infinitive in post-position, of which 149 are immediately followed by their infinitive. RQ6 asks about the *accentuation patterns* in these 149 configurations. The answer is visualized in Figure 6.

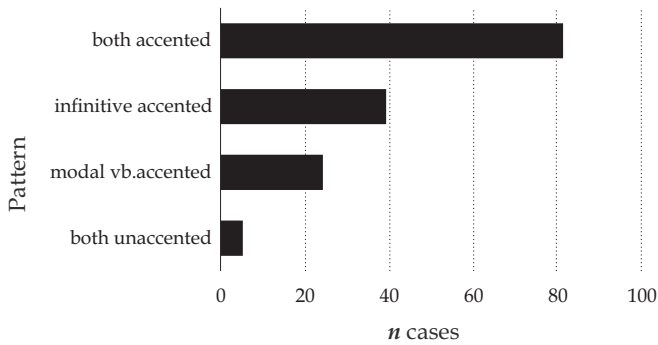


Figure 6 Patterns of accentuation in modals followed immediately by the infinitive.

The results resonate with the answers to previous RQs, but provide clearer numerical detail. In majority of the cases there, accents are on both the modal verb and the associated infinitive (n = 81, that is 54.4%). On the other hand, unaccenting of both elements occurred only five times (3.4%). If one of the verbs was unaccented, then it was the modal more often (second column from the top in Fig. 6).

Just to complete the description of this subject matter, we would like to add some information on the words separating the modal verb and the associated infinitive (per-

haps as a cue for future experimental design or cross-genre comparisons, etc.). As stated above, there were 76 cases of modal verbs followed by one or more other words before their associated infinitive. Table 3.2 shows how many times certain counts of intervening words occurred in the sample.

Table 3.2 Counts of words intervening between modals and associated infinitives in the corpus (1st line) and numbers of their respective occurrence in the sample (2nd line).

n intervening wds.	1	2	3	4	5	6
n times in the sample	34	19	12	4	6	1

Mostly, there was just one word between the modal and infinitive ($n = 34$), but there was also an unusual case of six intervening words. It is clear from (3) that this case involved an intervening clause (the modal and infinitive are underlined).

(3) *co lze bez obav, že bychom se mýlili, považovat za jedno z nejohavnějších období…*
what can be without fear that we would be mistaken denounced as one of the foulest periods…

As to syntactic or semantic functions of the intervening words, they were mostly adverbials (37 times), objects (10 times), or combinations of adverbial + object (5 times). Interestingly, a syntactic subject also occurred in the position between modal and its associated infinitive, but this happened only three times. The rest were particles, reflexive pronouns and various mixtures of synsemantics.

The final research question to be answered (RQ7) focuses on negative forms, which in Czech are constructed with the prefix *ne-* (e.g., *mohl = could* × *nemohl = could not*). Although it is possible to have a negative infinitive after a modal, or even simultaneous negation on both modal and infinitive, our sample did not contain any of such cases. There were only negative modal verbs with positive infinitives, and we found 54 of them. Only four of these were unaccented, which leaves 50 (92.6%) negative modals accented. This prevalence is commented on below, in the Discussion.

4. Discussion

More than two thirds of the modal verbs in the sample were accented, while about 30% were unaccented. That clearly differs from the situation both in autosemantic (full) verbs and in auxiliary verbs reported in Volín & Hanžlová (2024). However, the answer to RQ2 (*number of monosyllabic forms in accented and unaccented sets*) suggests that the communicative importance of modal verbs is not the singular force to determine their prominence, and that rhythmic (structural) factors exert quite a strong influence on the actual prosodic form. The need to avoid stress clash contributes to the fact that the monosyllabic forms of modal verbs remain unaccented much more often than the polysyllabic forms.

The examination of accentuation of infinitives associated with modals (RQ3) showed that the infinitives are prosodically backgrounded in only about 15% of the cases. That is just a half of the cases compared with the modals. This naturally leads to the question

of monosyllabicity, through which speech rhythm (more specifically the Stress-Clash Rule) manifests its power, and a subsequent comparison of the ratios of unaccented to accented cases in modals and in associated infinitives (RQ4). Monosyllabic modal verbs were found to be unaccented in the ratio of 2 : 1, monosyllabic infinitives only in the ratio of 1 : 1. When addressing the question of factors that cause the difference, one could hardly argue that modal verbs are more predictable or communicatively less important than their infinitives. Modals express very important subjective stances of the speakers to the pieces of reality that are being thematized. Therefore, we suggest that thanks to their high frequency of occurrence in speech, modal verbs are relatively easy to recognize and speakers intuitively save energy on them whenever the situation allows for it.

RQs 5 and 6 (*mutual position and accentuation pattern of modals and their infinitives*) also offer inspiration for further research, for instance, in the area of word order. In roughly a third of the cases, there was one or more words intervening between the modal and the associated infinitive. It would be interesting to know how such cases contribute to cerebral processing costs or, on the other hand, to the naturalness of the wording used. A perceptual experiment with pairs of utterances differing only in the distance between the modal and its infinitive would be useful, especially if realistic contexts of the utterances were involved.

RQ7 directed our attention to the modals in negative forms. Those were found hardly ever unaccented (only in 7.4% of the cases). There are two factors to consider when interpreting such finding. First of all, the negative prefix *ne-* makes the word form one syllable longer and it is a well-known fact that the increasing number of syllables in the word raises the probability of the accented spoken form (Volín & Skarnitzl, 2020). Second, the negative has a special semantic (or rather pragmatic) meaning – we could argue that positive forms are unmarked, while the negative ones are marked. Such pragmatic markedness could easily attract greater prosodic prominence.

There are many future tasks that stem from the current study. Importantly, there is the need to look into a finer classification of modal verbs (like, e.g., de la Rosa & Romero, 2021). For instance, it is possible that the *deontic* use of modals interacts with prosodic forms somehow differently from the *epistemic* use. Certain challenge in this area, though, would be the plethora of classification schemes offered by various scholars (see, e.g., the references in the fourth paragraph of Introduction). Nevertheless, van der Auwera and Plungian's classification (1998) seems suitable for Czech, and projects like these could actually test their suitability.

One line of follow-up projects should focus on cross-language or cross-genre comparisons. It is suggested that modality as such is a universal phenomenon, only expressed differently in different languages (Palmer, 2001; Nauze, 2008). That would extend the research to modal particles, modal adjectives, etc. One the other hand, even if we stick to modal verbs, various speech communication genres are very much likely to produce quite disparate patterns of modality expression. To go into further detail, one might be interested in sound change and see if older speakers treat modals differently from younger speakers or if there is any difference between recordings made, let us say fifty years ago and recently. That would naturally require a sample of audiobooks collected with attention to appropriate criteria.

Last but not least, the perceptual consequences of reversed accenting or unaccenting should be investigated. Those relate quite closely to the questions of effectiveness

in speech communication, and current behavioural or neuroimaging procedures could provide valuable findings if based on realistic speech material.

We hope that our study will be appreciated in the above-mentioned instances.

Acknowledgement

The study was carried out with the support of GAČR (Czech Science Foundation), Project 24-10905S "*Prosodic expression of utterance information structure in Czech*".

REFERENCES

Auwera, van der, J., & Plungian, V. A. (1998). Modality's semantic map. *Linguistic Typology, 2*(1), 79–124.

Benešová, E. (1973). K sémantické klasifikaci českých modálních sloves. In *Otázky slovanské syntaxe III*, Brno: Universita J. E. Purkyně, 217–219.

Boersma, P., & Weenink, D. (2022). *Praat: Doing Phonetics by Computer* (Version 6.2). [Computer Program]. Available online: www.praat.org (accessed in April 2022).

Daneš, F. (1979). O identifikaci známé (kontextově zapojené) informace v textu. *Slovo a slovesnost, 40*(4), 257–270.

Féry, C. (2017). *Intonation and Prosodic Structure.* Cambridge University Press.

Franz, I., Knoop, Ch., Kentner, G., Rothbart, S., Kegel, V., Vasilieva, J., Methner, S., Scharinger, M., & Menninghaus, W. (2022). *Prosodic Phrasing and Syllable Prominence in Spoken Prose. A Validated Coding Manual*. OSF Preprints. https://doi.org/10.31219/osf.io/h4sd5.

Geigerich, H. J. (1992). *English Phonology: An Introduction*. Cambridge University Press.

Grepl, M., & Karlík, P. (1986). *Skladba spisovné češtiny* [Syntax of Standard Czech] (1st ed.). Státní pedagogické nakladatelství.

Hengeveld, K. (2004). Illocution, mood and modality. In G. Booij, C. Lehmann, J. Mugdan, & S. Skopeteas (eds.), *Morphology: An International Handbook on Inflection and Word-Formation*, Vol. 2. de Gruyter.

Hoye, L. F. (2005). "You may think that; I couldn't possibly comment!" Modality studies: Contemporary research and future directions. Part I. *Journal of Pragmatics, 37*(8), 1295–1321.

Hualde, J. I. (2010). Secondary stress and stress clash in Spanish. In M. Ortega-Llebaria (ed.), *Selected Proceedings of the 4th Conference on Laboratory Approaches to Spanish Phonology*. Cascadilla Proceedings Project, 11–19.

Karlík, P., & Šimík, R. (2017). Modální sloveso. In Petr Karlík, Marek Nekula, & Jana Pleskalová (eds.), *CzechEncy – Nový encyklopedický slovník češtiny*.
URL: https://www.czechency.org/slovnik/MODÁLNÍ SLOVESO (retrieved: 02-02-2025).

Klinge, A. (1993). The English modal auxiliaries: from lexical semantics to utterance interpretation. *Journal of Linguistics, 29*(2), 315–357.

Lunden, A. (2019). Explaining word-final stress lapse. In R. Goedemans, J. Heinz, & H. van der Hulst (eds.), *The Study of Word Stress and Accent: Theories, Methods and Data*, 76–101. Cambridge University Press.

Nauze, F. (2008). *Modality in Typological Perspective*. Doctoral thesis DS-2008-08, University of Amsterdam.

Palková, Z. (1994). *Fonetika a fonologie češtiny* [Phonetics and phonology of Czech] (1st ed.). Univerzita Karlova, Nakladatelství Karolinum.

Palmer, F. R. (2001). *Mood and Modality*, Cambridge Textbooks in Linguistics (2nd ed.), Cambridge University Press.

Rosa, de la V. M., & Romero, E. D. (2021). Epistemic and non-epistemic modals: The key to interpreting the spirit of counter-terrorism United Nations Security Council resolutions. *Journal of Pragmatics, 180*, 89–101.

Svoboda, K. (1967). Dvě úvahy ke spisu K. Welkeho o „modálních" slovesech v němčině (O komunikačním efektu a o modálních slovesech). *Slovo a slovesnost, 28*(4), 426–431.

Volín, J., & Hanžlová, A. (2024). Deaccented verb as an element in the utterance information structure. In *Proceedings of Speech Prosody 2024*, 911–915. Leiden: ISCA. https://doi.org/10.21437/SpeechProsody.2024.
Volín, J., Šturm, P., Skarnitzl, R., & Bořil, T. (2024). *Prosodic Phrase in Spoken Czech*. Karolinum. https://doi.org/10.2307/jj.20558244.
Volín, J., & Skarnitzl, R. (2020). Accent-groups vs. stress-groups in Czech clear and conversational speech. In *Proceedings of Speech Prosody 2020*. Tokyo: ISCA, 695–699. https://doi.org/10.21437/SpeechProsody.2020-142.
Volín, J., & Skarnitzl, R. (2018). *Segmentální plán češtiny* [Segmental plan of the Czech Language]. FF UK.

APPENDIX

List of verbs that were analyzed as modal in the current study together with the number of occurrences in the sample (cf. Table 3.1 above).

General denotation	Verbs	total
obligation	muset (48), mít (34)	82
possibility	moct (50), jít/lze (3), dát se (1)	54
ability	moct (26) umět (12), dovést (5), dokázat (3)	46
permission	smět (6)	6
intention	chtít (36), snažit se (3), mínit (2), chystat (1), hodlat (1)	43
assigned obligation	nechat (3), dát (si) (2)	5

RESUMÉ

Tato studie se věnuje akcentování (realizaci přízvukového potenciálu) modálních sloves ve spojité řeči. Materiál pochází z audioknih, v nichž profesionální čeští herci (8 žen + 8 mužů) posloužili jako mluvčí, když četli vyprávění různých zavedených autorů. Zkoumaný vzorek obsahoval přes 17 000 slov, vždy nejméně 1 000 slov na mluvčího. Zvláštní pozornost při anotaci zvukových souborů byla věnována přítomnosti či absenci realizovaného slovního přízvuku (akcentu). Hlavním cílem této exploratorní studie je poskytnout referenční údaje týkající se modálních sloves v žánru narativního monologu. Výsledky ukazují, že modální slovesa se ve více než čtvrtině svých výskytů objevují bez realizovaného přízvuku, tedy neakcentovaná. Jednoslabičné formy se však v tomto ohledu chovají jinak než formy víceslabičné. Prošetřili jsme také infinitivy asociované s modálními slovesy. U nich docházelo k neakcentování (prozodickému upozadění) méně často a také chování monosylab probíhalo podle jiného vzorce. Studie dále poskytuje údaje o vzájemné pozici modálního slovesa a asociovaného infinitivu, a taktéž o záporných formách modálních sloves. Získané výsledky by měly být využity v následném výzkumu, například ohledně podílu komunikativních a formálních požadavků při prozodické strukturaci promluv nebo při zkoumání rozdílů mezi různými komunikativními žánry.

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

2025 ACTA UNIVERSITATIS CAROLINAE PAG. 97–115
PHILOLOGICA 3

A CONTRIBUTION TO THE STUDY OF SPEECH TEMPO AND PAUSE VARIABILITY IN TWO DIFFERENT SPEAKING STYLES

JITKA VEROŇKOVÁ

ABSTRACT
This study analyzes speech tempo and pause variability in two speaking styles: read-aloud news and semi-spontaneous self-introductions. Recordings of 10 non-professional Czech female speakers were examined for speech rate, articulation rate, and pauses. Results show that the read-aloud news were significantly faster, with a higher tempo and lower pause volume. While within-genre variability in the news was observed, the introductions exhibited greater inter-speaker variation, particularly in pausing. A speaker's tempo in one style did not strongly predict their tempo in the other. These findings underscore the role of speaking style in shaping the temporal characteristics of speech.

Keywords: Czech; speech rate; articulation rate; speech tempo; pauses; news reading; speech elocution

1. Introduction

Speech tempo has long been a subject of interest to researchers. The sound features that cause listeners to perceive speakers as fast or slow, along with the factors influencing our perception of speech tempo[1], have been examined. Several factors, such as articulation rate, may have a greater impact; however, they do not act independently. The perception of speech tempo is affected by multiple interacting factors with considerable overlaps (cf. Kohler, 1986; Koreman, 2006; Plug et al., 2022).

Factors that have been examined for their influence on speech tempo production include, for instance, age, gender, and dialect region (e.g., Verhoeven et al., 2004; Quené, 2005; Yuan et al., 2006; Jacewicz & Fox, 2010; Bóna, 2014; Huszár & Krepsz, 2021; Ferguson et al., 2024). The results of research are not always consistent, and it is evident that the factors interact in complex ways.

The type of speech task is another factor that has been monitored. For example, Barik (1977) found that articulation rate, speech rate, and pausing were influenced by the speaker's degree of readiness during a given performance. Along with the role of the speech task or speaking style, inter-speaker and intra-speaker variability (or, conversely,

[1] In this paper, the term *speech tempo* (or simply *tempo*) is used as a general term.

https://doi.org/10.14712/24646830.2025.23

stability) has been monitored (e.g., Mixdorff et al., 2005; Jacewicz & Fox, 2010; Bóna, 2014; Huszár & Krepsz, 2021; Ferguson et al., 2024), usually by comparing read and (semi)spontaneous speech.

From an alternative perspective, Koopmans-van Beinum and van Donzel (1996) proposed speech rate to be one of the two main cues listeners use to differentiate between spontaneous and read speech. Their research on spontaneous speech supported the belief that variation in speech rate is somehow related to the information structure in the discourse.

In the first decade of the 21st century, a couple of contributions to the topic of speech tempo in Czech were published (Dankovičová, 2001; Veroňková-Janíková, 2004; Balkó, 2005). They all used recordings of non-professional speakers in various speech tasks. Dankovičová, who focused on articulation rate, suggests that recurring temporal patterns appear within a prosodic unit (Dankovičová, 2001). The speech material analyzed by Balkó encompassed several speaking styles, and the results demonstrated that both articulation and speech rates are influenced by the type of task and the degree of difficulty in planning the speech; at the same time, individual differences exist among speakers (Balkó, 2005). Similar results were reported by Veroňková (2004) based on six different tasks, including read-aloud vs. semi-spontaneous speech, contrasted by recording environment (i.e., individual recording in a studio vs. a semi-public performance). For example, narrating a fairy tale based on a series of pictures proved to be the slowest task by far. There was a clear tendency for a higher speech rate to be associated with the read texts. However, some speakers decreased their speech rate in the read version of a story compared to the original spoken one. This could be explained by an intentional effort to slow down, as the original tempo of the story was criticized by the audience as being too fast.

The present paper was inspired, among others, by two studies by Volín (2019, 2022), published in this AUC Philologica series, both in terms of their topics and methodology. Volín (2022) examined temporal differences in two genres – news reading and poetry reciting – and provided reference values for some tempo metrics. Two target genres differed in both articulation and speech rates (news reading was faster than poetry); however, the results suggested that articulation rate was more stable in both between-genre and within-speaker comparisons, meaning that pausing was more variable. Temporal characteristics between genres showed mostly individual speaker behaviour, while inter-speaker variability within a genre was low, which, according to Volín, might suggest shared communication concepts.

The objective of the present study is to provide data concerning the variability of speech tempo and pauses across two distinct speaking styles. The perspective is motivated by a long-term ambition to examine the behaviour of speakers in various speech situations and to analyze speech performances as a whole, including the impact on listeners. The speech material for the present study was provided by a sample of non-professional speakers, each of whom produced speech in both targeted speaking styles. The first style is news reading, which exemplifies the so-called 'clear speech'. The second is a semi-public speech performance delivered in front of an audience; in the recordings analyzed, this style closely resembles ordinary conversational settings (for details, see the Method section).

It is precisely the inclusion of a semi-public performance that makes the current study unique. To the author's knowledge, such speech material has not been used in recent

studies. Recordings of (semi-)spontaneous speech have been analyzed, but these typically consisted of either speech in the presence of an interviewer or recordings of individuals in either pairs or very limited groups. A side benefit of the current analysis was also checking the technical quality of these semi-public recordings and the possibility of their use for further phonetic experiments.

Given the nature of the phenomenon under investigation, the target speaking styles and genres represent an appropriate choice. Non-professional speakers may use speech tempo to emulate professional news readers, given the general assumption – supported by objective measurements – that news readers use a higher speech tempo (Veroňková & Poukarová, 2017). Speech tempo is also one of the factors addressed by rhetoric and is often the subject of evaluation of a speaker's performance, particularly in relation to pausing. Pauses are a noticeable phenomenon for the listener and a very clear signal of a prosodic unit boundary. This is why the inter-pause stretch was taken as the basic unit for measurement in this phase of research. Although articulation rate exhibits regular patterns within a prosodic unit (for Czech, see Dankovičová, 2001), this study focuses on this larger interval. With regard to speech tempo and pausing, the paper aims to examine between- and within-genre variability and inter-speaker variability.

2. Method

2.1 Material

2.1.1 Speaker and recording selection

The recordings were sourced from an archive of student recordings created during mandatory courses in the phonetics programme at the Faculty of Arts, Charles University. None of the recordings were made specifically for the purposes of the research presented in this paper. The author conducted all recordings.

For this study, recordings of 10 female speakers with Czech as their mother tongue were used. Their ages ranged from 19 to 23 years, and they did not report any neurological, speech, or hearing disorders. Two recordings of two speech genres were used from each speaker: a semi-spontaneous self-introduction and read-aloud news.

The speakers and their recordings were selected from the archive using the following procedure. Students from eight phonetics courses who recorded the same news bulletin text formed the baseline group of speakers. From this group, the following speakers were excluded: a) non-native speakers of Czech, b) men – due to their limited representation in the archive, and c) recordings by females that contained a large number of errors and slips of tongue. These recordings were omitted to ensure comparable content across all speakers, as dysfluent passages would have otherwise been removed during the measurement phase. The author then matched the selected news reading recordings with the self-introduction recordings from the same speakers. A few pairs were discarded due to the low technical quality of the introduction recording. Conversely, the degree of dysfluency in the self-introduction did not play a role in the selection process.

2.1.2 Speech material and recording environment

Self-introduction speech

The recording of the self-introduction took place in groups of approximately 12–15 people at the beginning of the Speech elocution course. The speakers' task was to introduce themselves. They were given a list of points to talk about (name, age, high school, university studies, etc.), but were free to decide whether to cover all of them and how to structure their presentation. The speech was intended to last approximately 1 minute. Speakers were given approximately 5 to 10 minutes to prepare before taking turns individually. They were allowed to make notes during preparation; however, they performed without them. The target speaker delivered the speech using a hand-held microphone in front of the others, who served as the audience. The talk was also recorded on a video camera.[2] The entire session, including the preparation, lasted approximately 60 minutes.

With the exception of two respondents, the speakers gave their self-introduction presentation in the first semester of their university studies, often during their very first class, i.e., in an unfamiliar group. They were informed – without further specification – that the recording would provide material for course work and serve as an opportunity for them to get to know each other. The speeches were affected by varying degrees of nervousness, evident from both the recordings and the speakers' accompanying comments. The teacher (the author) attempted to create a friendly atmosphere to reduce tension and encourage the speakers.

The recordings were obtained in a medium-sized classroom (approx. 30 seats) with soundproofed walls. The hand-held microphones (Sennheiser e840 or Rode NT3) were plugged into a computer sound card and recorded using Praat (Boersma & Weenink, 2016) in WAV format with 22.05-kHz sampling frequency using a 16-bit resolution, or using Audacity (48 kHz, 16-bit). Although some recordings were of slightly lower quality, high-quality recording was not essential for analyzing the temporal phenomena under investigation.

News bulletin

The speakers read the same *bulletin* of six *paragraphs* (approx. 550 words in total), and the entire text was included in the analysis. The paragraphs are not completely balanced in terms of syllable count. The first and last paragraphs contain introductory and concluding phrases, respectively (they were also included in the analyses and referred to as *N_ini* and *N_fin*). The news reading genre itself is represented by four paragraphs (*N1–N4*), which are shortened versions of real news from the national broadcaster, Czech Radio[3]. The bulletin text does not contain any lexical items with unclear pronunciation, such as foreign words or names. Apart from the shortening of individual news items, no other modifications were made to the text.

The bulletin recording was part of a session that students completed (with two exceptions) at the beginning of their second year in the phonetics programme, i.e., a year after the self-introduction recording. The news reading was performed individually in the

[2] It does not apply to two speakers from our list.
[3] Český rozhlas, https://portal.rozhlas.cz.

sound-treated studio of the Institute of Phonetics in Prague. The session typically consisted of two rounds, with respondents successively recording the news and then another read text of a different genre (or vice versa). Speakers were given a paper copy of the target text and time to familiarize themselves with it prior the reading. They were allowed to make notes on the sheet and take it into the studio. No special instructions concerning the reading style were provided.

2.2 Procedure

2.2.1 Data segmentation and annotation

The sound processing was performed in Praat (Boersma & Weenink, 2016). Phone and word boundaries (based on orthographic transcripts from archive) were forced-aligned using the Prague Labeller (Volín et al., 2005; Pollák et al., 2007) and Prak (Hanžl & Hanžlová, 2022, 2023). The boundaries were then manually corrected, with special attention to pauses, following the annotation rules presented in Machač & Skarnitzl (2009). For the purposes of this study, prepausal vowels were labelled with an emphasis on perception rather than formant structure.

Pauses were defined as sections without lexical articulation and could be either silent or filled. The boundaries of pauses were labelled manually based on the author's perception; therefore, no minimal pause duration was determined. The vast majority of boundaries were clearly located. Discretion was required in a limited number of cases, for example, when strong glottalization was present or when a canonical word-initial/ /final sound was preceded/followed by an additional schwa. In such cases, the sound was considered part of the articulation if it appeared to be an integral component of the word. If it gave the impression of a hesitation, it was labelled as part of a pause. This process determined the boundaries for the *inter-pause stretches*.

In the bulletin recording, pauses were mainly filled with breaths. In the self-introduction, pauses also contained noticeable hesitation sounds, especially for some speakers. To check whether this could be a source of variability, hesitation passages within the pauses were also separately labelled in these two performances.

One speaker's introduction was twice as long as the others; therefore, only less than the first half of this speech was used for measurement; the cut was made at the point of semantic completion where the speaker finished a subtopic. In one self-introduction recording, a couple of shorter inter-pause streches that the speaker delivered with laughter were excluded.

Due to the speaker selection criteria, the bulletin recordings did not contain significant dysfluencies requiring exclusion. There were a few minor slips of tongue (9 in total, 0–3 per speaker), which were always located in the middle of an inter-pause stretch and did not disrupt the speech flow[4].

The resulting blocks of speech are referred to as *performances*. To obtain units parallel to the bulletin paragraphs, the self-introduction was divided into four parts –

[4] For example: *V České republice se [ro] loni narodilo nejvíce dětí…* Eng: *The Czech Republic had the highest number of children born last year…*

an initial part, two middle parts, and a final part (*P1–P4*). The self-introductions typically contained four subtopics, which made these divisions relatively easy. Not all self-introduction performances contained final phrase (such as *Thank you for your attention*), so this phrase was not included in the final (P4) paragraph but is referred to as *P5* instead, if present.

2.2.2 Measurements and analysis

To determine the number of syllables and calculate the articulation rate, scripts developed by Oceláková and Bořil were used (Oceláková, n.d.; Bořil & Oceláková, n.d.). The data processing and subsequent analysis were partially performed in R within RStudio environment (R Core Team, 2024). Statistical tests were performed using the online calculator at Statistics Kingdom (2017a, 2017b). One-way ANOVA for repeated measures, the post-hoc Tukey HSD test, and Pearson correlation were used to test significance of the results. Outcomes of the statistical tests were considered significant at an alfa level of 0.05.

Adopting the methodology of Volín (2022), this study examined several descriptors of variation: minimum, maximum, their distance (i.e., variation range; all in syll/s), and the coefficient of variation (Cvar, in %), calculated as the ratio of the standard deviation to the arithmetic mean, multiplied by 100. The interpretation of the Cvar values is based on the thresholds presented in Volín (2022): Cvar < 30% represents concentrated values, while Cvar > 50% represents dispersed values.

Pause volume (in %) and pause duration (in seconds) were also monitored. Duration values were used in both non-normalized and normalized forms. Normalization was performed using the following formula: durnorm = dur * ARindiv / ARspeakers (where ARindiv is the AR of a given speaker in a given performance, and ARspeakers is the mean AR of all speakers in that performance). The same descriptors of variation used for tempo were applied to pauses.

A limit value was determined for assessing changes in temporal course; according to Quené (2007), the just-noticeable difference is approximately 5%, which corresponded to 0.3 syll/s for both SR and AR in our speech material.

The analysis employed two analytical units: *performances* and *paragraphs*, with the latter corresponding to the *genre unit* used by Volín (2022). As the long-term focus of this research is on speakers' presentations as a whole and their impact on listeners, the primary analysis encompassed speech stretches of varying lengths, treating them as integral components of the overall performance. In the next step, the data were also recalculated, first excluding 1-syllable stretches and then excluding 1-syllable stretches together with 1-word stretches.

Two sources of data were used: a) *overall values* (the total duration and corresponding number of syllables for a performance/paragraph), and b) the *articulation rate of single inter-pause stretches*. The ARs of individual stretches were used to calculate a weighted mean for a given unit (weighting was done by duration).

The study focuses on two types of performance – news bulletin reading and self-introduction – each produced by ten speakers. Two speaking styles were examined: read-aloud news and semi-public speech presentation.

3. Results

The results will be presented in the following order: (1) the differences between genres regarding speech tempo and pauses, (2) within-genre differences regarding speech tempo, (3) inter-speaker differences. Self-introduction is further referred to as *introduction* or *intro*.

3.1 Between-genre differences

3.1.1 Rates in performances

The measurements depicted in Figure 1 are based on overall performance values of the bulletin and the introduction.

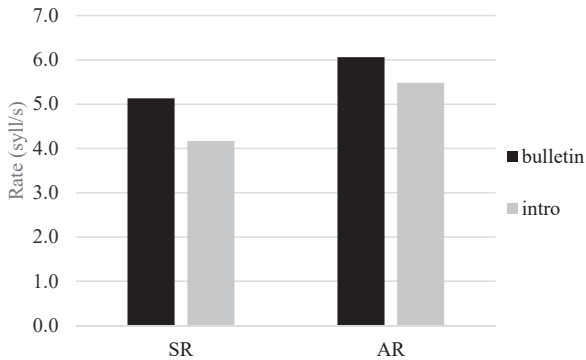


Figure 1 Mean speech and articulation rates in two performances: bulletin (N = 10), and introduction (N = 10). The results are based on overall values.

Although the study focuses on the difference between genres, the relationship between SR and AR within the same genre will be examined first. Given the calculation of SR and AR differs in pauses, which are present in both performances, we assumed that SR and AR would differ. As expected, SR and AR differed within both the bulletin and introduction performances (compare the left and right sides of each pair in Figure 1). A one-way ANOVA for repeated measures returned strong significant results for both performances. Bulletin, $F(1, 9) = 329.7$, $p < 0.001$; introduction, $F(1, 9) = 227.4$, $p < 0.001$.

Regarding between-genre differences, both SR and AR differed between the bulletin and introduction (compare SRs and ARs separately in Figure 1). The bulletin reading was faster, with an SR higher by 1.0 syll/s and an AR higher by 0.6 syll/s. A one-way ANOVA confirmed a significant difference for both rates, although the effect was slightly weaker for the articulation rate. SR: $F(1, 9) = 42.9$, $p < 0.001$; AR: $F(1, 9) = 19.6$, $p \approx 0.002$.

In addition to using overall values, AR was also calculated as a weighted mean of the AR from individual inter-pause stretches. A one-way ANOVA confirmed a significant difference in these recalculated ARs between the bulletin and intro as well. $F(1, 9) = 18.6$, $p \approx 0.002$.

The computation of variation metrics (Table 1) was based on 60 data points for the bulletin (10 speakers × 6 paragraphs) and 40 data points for the introduction (10 speakers

× 4 paragraphs). This means that transition pauses between paragraphs and any potential final phrases (P5) in the introduction were excluded from this calculation.

Table 1 Variation metrics for speech rate (SR) and articulation rate (AR) across the bulletin and introduction performances, expressed in syllables per second (syll/s).

Rate – genre	C_{var} (%)	Min	Max	Range
SR – bulletin	8.9	4.4	6.9	2.5
SR – intro	16.0	3.2	6.2	3.0
AR – bulletin	6.8	5.3	7.1	1.7
AR – intro	11.2	4.3	6.7	2.4

The values for the bulletin are more compact than those for the intro, for both SR and AR. The minimum rates observed in the bulletin were higher than those recorded in the intro (by 1.2 syll/s for SR and 1.0 syll/s for AR). The maximum rates recorded in the bulletin were also higher, although the difference was less pronounced (0.7 syll/s for SR and 0.4 syll/s for AR). Consequently, the range is narrower for the bulletin for both rates.

The coefficient of variation (Cvar) for bulletin was below 10%, indicating highly concentrated values. The Cvar for the intro was somewhat higher, particularly for SR (16.0%), while the AR value of the intro (11.2%) only slightly exceeded 10%.

3.1.2 Pauses in performances

The measurements depicted in Figure 2 are based on overall performance values.

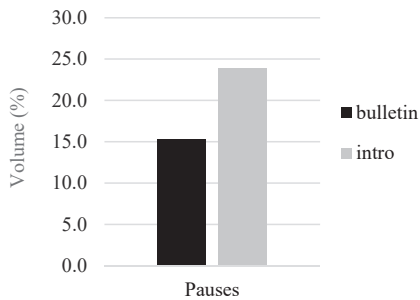


Figure 2 Mean volume of pauses (as a percentage of total duration) in two performances: bulletin (N = 10), and introduction (N = 10). The results are based on overall values.

It is clear that the genres differ in their overall pauses volume; in the bulletin, pauses constitute, on average, approximately 15% of the performance, while in the intro, they account for about a quarter of the duration. An ANOVA for repeated measures returned a significant difference between the bulletin and intro, $F(1, 9) = 29.3$, $p < 0.001$. A higher Cvar value was measured for the intro (25.1%) than for the bulletin (16.0%); however, both values still indicate relatively compact data.

In addition to pause volume, pause duration was also measured. Figure 3 depicts the mean pause duration in two forms: as measured (non-normalized) data and normalized

data. The contrast between these two displays is evident: while the measured mean pause duration reveals differences between bulletin and introduction (left side of the figure), the normalized pause durations remain relatively consistent across both performance types (right side). This observation is confirmed by statistical significance testing. A one-way ANOVA returned a non-significant result for the normalized data, $F(1, 9) = 0.1$, $p \approx 0.7$. However, for the non-normalized pause duration, the result is significant, $F(1, 9) = 10.6$, $p \approx 0.010$.

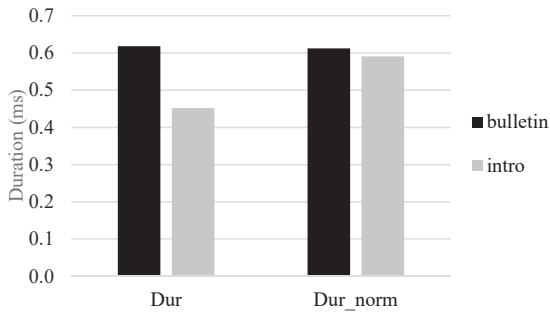


Figure 3 Mean pause duration (non-normalized and normalized) in two performances: bulletin (N = 10), and introduction (N = 10). The results are based on overall values.

It must be noted that although the total volume of pauses in the bulletin is significantly lower than in the intro, the mean non-normalized duration of a single pause is higher (and vice versa for the intro).

3.2 Within-genre differences

This analysis is based on paragraph-level data, which excludes transition pauses. For the bulletin, six paragraphs were examined, and for the introduction, four paragraphs were analyzed. Each paragraph represents one *genre-unit* (cf. Volín, 2022).

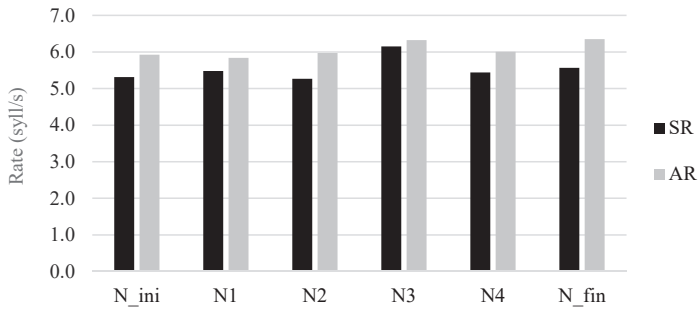


Figure 4 Mean speech and articulation rates in bulletin, divided into six paragraphs. (N = 10 for each paragraph)

3.2.1 Bulletin

Figure 4 shows that the tempi of the paragraphs are very similar, except for N3, where the SR and AR do not differ to such extent from each other. An ANOVA test for repeated

measures indicated that there is a significant difference in SR among the news paragraphs $F(5, 45) = 6.53$, $p < 0.001$ ($\alpha = 0.05$). The post-hoc paired t-test using a Bonferroni corrected $\alpha = 0.0033$ indicated that the means of the following three pairs are significantly different: N3 on the one hand, and N1, N4 and N_fin on the other.

The difference between paragraphs is also significant for AR as returned by an ANOVA for repeated measures: $F(5, 45) = 17.28$, $p < 0.001$ ($\alpha = 0.05$). The post-hoc paired t-test using a Bonferroni corrected $\alpha = 0.0033$ indicated that the means of the following eight pairs are significantly different: N3 and N_fin on the one hand, and N_ini, N1, N2, N4 on the other.

The correlation between the SR and AR values for the paragraphs was also tested to examine whether higher SR scores tend to co-occur with higher AR scores. The Pearson's $r(58)$ value is 0.838, $p < 0.001$. This represents a strong positive correlation, statistically significant.

As noted previously, the paragraph-level analysis excludes transition pauses. Within the bulletin, a noticeable difference was attested between the duration of transition pauses (pauses between paragraphs) and inner pauses (within paragraphs). The mean non-normalized duration was 1.41 s for transition pauses (SD = 0.28 s) and 0.43 s for inner pauses (SD = 0.09 s). The lower Cvar for transition pauses suggests greater uniformity in their duration. However, at nearly 30%, this value approaches the upper limit of what is considered compact data. In contrast, inner pauses exhibited the Cvar of almost 60%, indicating highly dispersed data.

The question arises as to what happens when the distinction between transition and inner pauses in the bulletin is considered comparing pause durations between the bulletin and the intro (3 types of pauses x 10 speakers = 30 datapoints). A one-way ANOVA returned a highly significant result: $F(2, 18) = 123.8$, $p < 0.001$. A post-hoc paired t-test using a Bonferroni corrected $\alpha = 0.01667$ indicated significant differences between the transition pauses in the bulletin on the one hand, and the inner pauses in the bulletin and the pauses in the intro on the other; the difference between the inner pauses in the bulletin and the pauses in the intro was not significant.

3.2.2 Introduction

Figure 5 shows the SR and AR values for the introduction divided into four paragraphs.

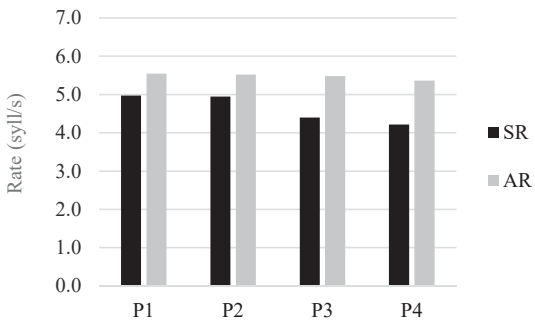


Figure 5 Mean speech and articulation rates for the introduction divided into four paragraphs. (N = 10)

The relationships among the paragraphs of the introduction were examined. A one-way ANOVA returned a significant result for the SR of paragraphs P1–P4 (4 paragraphs x 10 speakers), $F(3, 27) = 4.1$, $p \approx 0.017$. However, the multiple comparisons did not identify a significant difference between any of the pairs. For AR, the lack of significant difference is apparent from the display and was confirmed by an ANOVA test: $F(3, 27) = 0.3$, $p \approx 0.80$.

3.2.3 Course of tempi in the bulletin

The similar tempo patterns observed across speakers suggest the text itself influences the delivery. (These data also relate to inter-speaker variability, see Section 3.3.)

Figure 6 depicting the tempo course between paragraphs with respect to individual speakers shows a conspicuous pattern for SR, in particular in the last four paragraphs. A clear trend may be observed: an increase from N2 to N3, followed by a decrease to N4. The increase to N3 is evident in all speakers (the increase exceeds even 1.0 syll/s in three speakers). The subsequent decrease to N4 is not as extensive and occurs for nine of the ten speakers. For AR, this trend may also be noted, though less prominently (Figure 7).

Interestingly, the three speakers who did not increase their AR from N2 to N3 were among the fastest speakers overall. This observation leads to the assumption that speakers with a higher baseline articulation rate might exhibit less tempo variability. To test this, a Pearson correlation between each speaker's overall AR and their coefficient of variation (based on the inter-pause stretches) was calculated. The result shows only a weak, non-significant negative relationship ($r = -0.16$, $p \approx 0.65$). Furthermore, the Cvar values themselves are consistent across the group, falling within a narrow range of 9.6% to 13.9%, which indicates a similar level of tempo stability among all speakers.

For both SR and AR, the majority of speakers exhibited an increase in tempo between N4 and the final paragraph N_fin. In the case of SR, however, the extent of this increase was notably smaller than the rise observed between paragraphs N2 and N3.

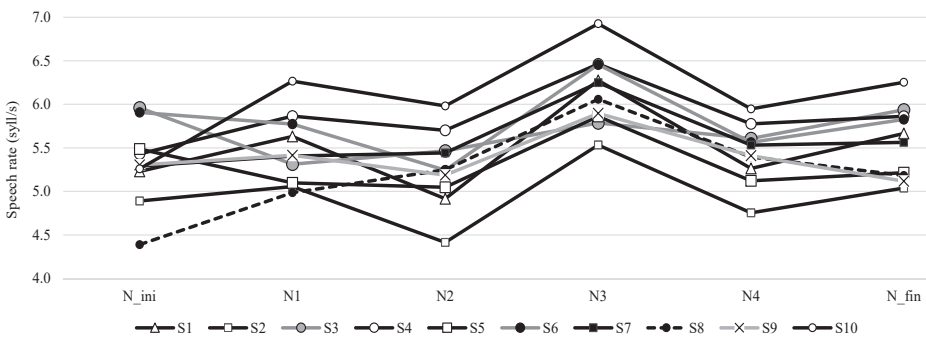


Figure 6 Mean speech rate in bulletin paragraphs, indicating the tempo course for individual speakers (S1–S10).

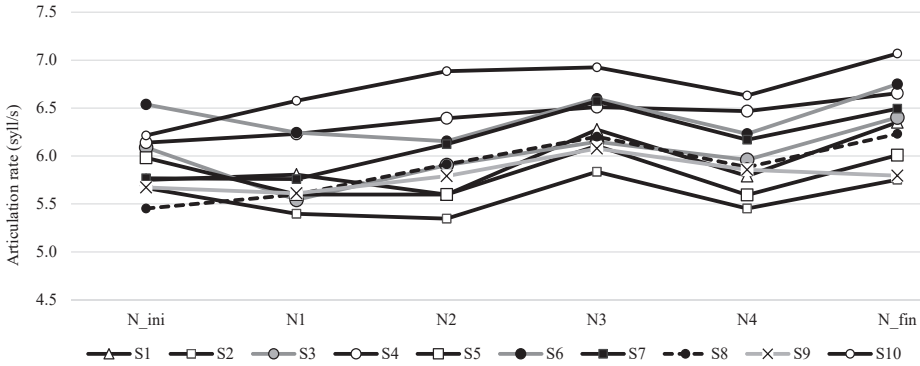


Figure 7 Mean articulation rate in bulletin paragraphs, indicating the tempo course for individual speakers (S1–S10).

3.3 Inter-speaker differences

3.3.1 Speech tempo

The displays in Figure 8 (bulletin) and 9 (introduction) are based on the overall SR and AR for each speaker.

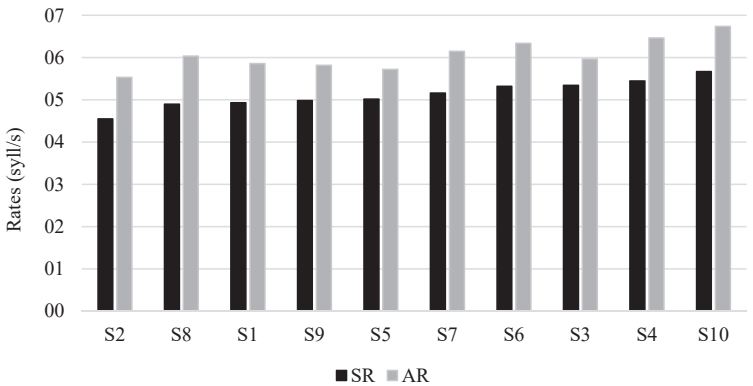


Figure 8 Mean speech and articulation rates produced by individual speakers (S1–S10) in the bulletin, ordered by SR values.

Across both genres, all speakers demonstrated a clear difference between SR and AR. Interestingly, within both the bulletin and the intro, the same speaker produced the minimum SR and AR, while another single speaker produced both the maxima in the bulletin; similarly, both maxima were provided by a single speaker in the introduction as well. According to the Pearson correlation test, SR and AR are strongly and positively correlated within both genres, meaning that high values of AR correlate with high values SR (and vice versa): $r = 0.90$, $p < 0.001$ (bulletin), $r = 0.87$, $p < 0.001$ (introduction).

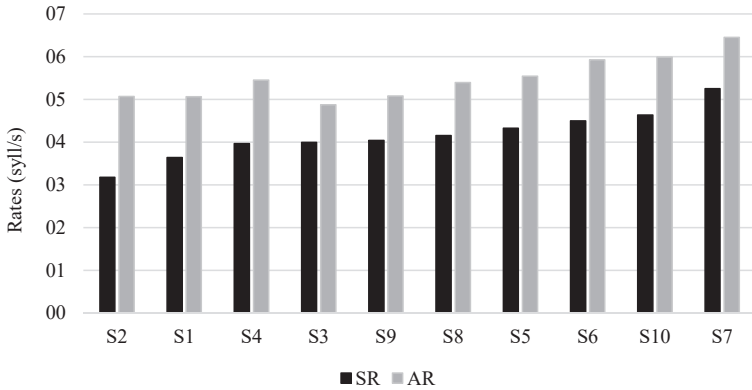


Figure 9 Mean speech and articulation rates produced by individual speakers (S1–S10) in the introduction, ordered by SR values.

A high degree of similarity among speakers within a single genre is evident, particularly in the bulletin. This observation is confirmed by the data presented in Table 2, which demonstrate low dispersion as evidenced by the Cvar values. Of the measurements, the only Cvar value to exceed 10% was for SR in the introduction. This higher variability corresponds directly to the introduction, having the widest SR range (maximum – minimum). Furthermore, the data show that the minima and maxima for both SR and AR were higher in the bulletin than in the introduction, indicating that the overall tempo was faster in the bulletin genre.

Table 2 Variation metrics for speech rate (SR) and articulation rate (AR), based on the individual speaker values for each performance. The rates are expressed in syllables per second.

Rate – genre	C_{var} (%)	Min	Max	Range
SR – bulletin	6.3	4.5	5.7	1.1
SR – introduction	13.6	3.2	5.3	2.1
AR – bulletin	6.1	5.5	6.7	1.2
AR – introduction	9.2	4.9	6.4	1.6

An examination of the ordering of individual SRs by magnitude reveals that certain speakers retained their relative positions across both genres (e.g., S2, see comments to minima above); some speakers moved their positions slightly (e.g., S1 and S10), while for others the change was greater (e.g., S8 and S4 moved by four and six positions, respectively). The question naturally arises as to how strong the correlation is between the speaker ordering across the two genres. A Pearson correlation test returned only a moderate, non-significant positive correlation. SR: $r = 0.57$, $p = 0.087$; AR: $r = 0.59$, $p = 0.073$.

To test the influence of very short stretches, AR for each speaker was calculated under three conditions. The first, set 1 (used so far), included all speech stretches regardless of their size. For set 2, 1-syllable stretches were excluded, and for set 3, both 1-syllable and 1-word stretches were excluded. An ANOVA for repeated measures indicated a signif-

icant overall difference in weighted mean of ARs among the three sets: $F(2, 18) = 23.4$, $p < 0.001$. The post-hoc paired t-test using a Bonferroni corrected $\alpha = 0.01667$ indicated that the means of all three pairs differ significantly. However, the magnitude of the difference between the averages is small. The largest difference observed between set 1 and set 3 was just 0.4 syll/s for one speaker and 0.3 syll/s for three others. For the majority (six speakers), no difference in the calculated AR was noted. These results suggest that while the inclusion of short stretches produces a statistically significant effect overall, its practical impact on AR values is negligible for most speakers.

3.3.2 Volume of pauses

Figure 10 displays the volume of pauses for individual speakers based on overall values.

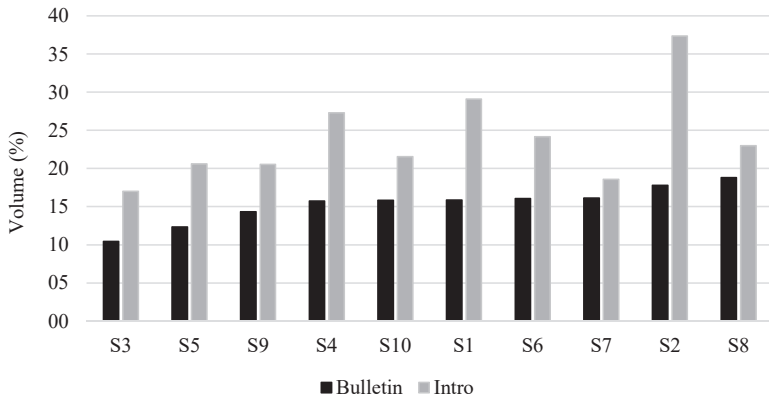


Figure 10 Volume of pauses (in %) produced by individual speakers (S1–S10) in the bulletin and introduction, ordered by percentage in the bulletin.

In contrast to the bulletin, inter-speaker differences in the volume of pauses in the introduction are evident. For example, speaker S7 exhibits a notably low value in the introduction, comparable with the value in the bulletin. In contrast, three other speakers display values up to twice as high in the introduction as those recorded in the bulletin. It is clear that the values of pause volume between both genres are not correlated. A Pearson correlation test returned only a moderate, non-significant positive correlation: $r = 0.57$, $p = 0.083$.

The greater compactness of pause volume in the bulletin compared to the introduction is also evident from the coefficients of variation (15.9% vs. 24.9%) and the range (8.4% vs. 20.2%), as shown in Table 3.

Table 3 Variation metrics for pause volume (in %), based on individual speaker values.

Genre	C_{var}	Min	Max	Range
Bulletin	15.9	10.5	18.8	8.4
Intro	24.9	17.0	37.2	20.2

The speech of some speakers in the intro contained perceptually noticeable passages with hesitation sounds. Figure 11 shows the volume of pauses for individual speakers in the introduction, indicating the ratio between parts without hesitation and those containing hesitation. The volume of hesitation sounds is a source of inter-speaker variability; several speakers' pauses were not filled with hesitation sound at all (S4) or very little (S7, S6). On the contrary, in some speakers the volume of these pauses was relatively high (S2, S5).

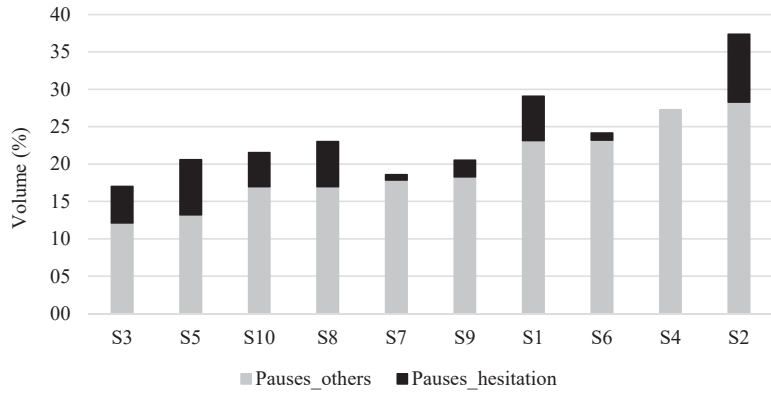


Figure 11 Volume of pauses (in %) produced by individual speakers (S1–S10) in the introduction, indicating the ratio between parts with and without hesitation (ordered by the magnitude of volume without hesitation).

There appears to be no correlation between the total volume of pauses and the volume of pauses containing hesitation sounds (compare, e.g., the display for speakers S5, S4, and S2). This is confirmed by a Pearson correlation test, which returned a non-significant result: $r = -0.02$, $p = 0.95$.

4. Discussion

The focus of the current study was on tempo variation and pausing in two distinct speech genres: a read-aloud bulletin and a semi-spontaneous self-introduction. The main outcomes are discussed in this section and compared with relevant findings of Volín (2022), who analyzed news reading and poetry reciting.

The two genres examined in the current study differed significantly in tempo. The read-aloud news bulletin was significantly faster than the self-introduction, in terms of both speech rate (SR) and articulation rate (AR). These outcomes are consistent with those of Volín (2022), where news reading was faster than poetry reciting. However, the differences between SR and AR observed in Volín's study were substantially greater than those identified in our dataset. This suggests that tempo of a semi-public introduction is closer to news reading than poetry reciting is.

The tempo in the bulletin was also much more compact than in the introduction, as indicated by variation metrics. The coefficient of variation (Cvar) values for both rates in the bulletin did not exceed 9%, which aligns perfectly with the corresponding value for

news in Volín's material. However, a difference emerged: while Volín found that only the speech rate in poetry showed high variation (with articulation rate remaining stable), the current study found that both rates (SR and AR) displayed higher variation in the introduction. Moreover, the Cvar for speech rate in the introduction was somewhat higher than that for poetry reading in Volín's sample (16.0% vs. 12.3%).

Regarding pauses, two aspects were examined: volume and duration. The introduction contained a significantly higher volume of pauses than the bulletin. Although the total pause time was lower, the average duration of individual pauses was significantly greater in the bulletin. When pause duration was normalized to account for each speaker's articulation rate, this difference between the genres became non-significant.

However, these overall data on pause duration masked an important finding. Transition pauses (between paragraphs) in the bulletin were significantly longer than inner pauses (within paragraphs). While the Cvar for transition pauses was close to the limit of compactness, the inner pauses were highly dispersed. Crucially, a between-genre comparison revealed that the inner pauses in the bulletin were not significantly different in duration from the pauses found in the introduction. These outcomes indicate that pauses in the bulletin likely serve the function of structural markers, with longer pauses at paragraph boundaries signalling a topic shift. In the future, not only overall pause volume but also the specific placement, frequency, and acoustic features of pauses should be examined in more detail, including their role in expressing information structure and their perceptual impact on listeners.

Regarding the analyzed speech material, it is important to consider an additional factor that may have influenced the observed genre differences, namely the time interval between the recordings, which was one year for most speakers. The examination of intra-genre differences revealed that certain news paragraphs differed significantly in tempo from others. Moreover, most speakers showed a similar pattern of tempo changes across the paragraphs when reading the bulletin. A preliminary examination suggests that factors such as the number of syllables or sentences may affect the tempo within individual news paragraphs. Volín (2022: 71) noted a mean tempo deceleration in news reading across paragraphs; in this context, he indicated (without further details) the potential influence of the content of the paragraph on tempo (e.g. domestic news vs. foreign news vs. sport). The extent to which linguistic content modulates speech tempo represents a promising area for future research.

In contrast to the bulletin, the introduction did not exhibit a consistent tempo pattern across the paragraphs. While some variation is attested, no single paragraph was consistently faster or slower than another. The articulation rate seemed stable across the paragraphs of the introduction.

Concerning inter-speaker differences, a speaker's SR and AR were strongly and positively correlated within each genre. However, a cross-genre comparison indicated that a speaker's tempo in the bulletin is not significantly correlated with their tempo in the introduction. This suggests that speakers may apply different tempo strategies in different genres, a finding consistent with Volín (2022), who also reported only a moderate correlation between speakers' performances in poetry and news reading.

Regarding within-genre variation, speakers displayed a high degree of similarity, particularly when reading the bulletin. Greater individual differences were noted in

the introduction, especially in speech rate. This aligns with Volín's observation of low inter-speaker differences within a given genre.

Large differences between speakers in the volume of pauses were evident only in the introduction; this measure was not correlated between genres, however. Inter-speaker variability in the volume of hesitation sounds was also high in the introduction; nevertheless, there was no correlation between a speaker's total pause volume and their use of hesitation. For example, the speaker with the second-highest overall pause volume produced no hesitations. This suggests that speakers have different strategies for planning speech or masking nervousness; some accumulate pause time through silent pauses (which may be either long or frequent), while others use filled pauses. This could be a fruitful area for future research, including examination of the impact of speech fluency on listeners.

The use of semi-public performances extends the range of speech materials often used in phonetic research. The presence of an audience likely heightened cognitive load and nervousness, thereby influencing speakers' pause behaviour. Compared to news reading, the self-introduction is a less constrained task where individual planning abilities become more prominent.

Although a relatively small sample was analyzed, the results of the current study suggest that speaking style and/or speech genre is not just a minor variable but a force that has the potential to reshape a speaker's entire temporal strategy, influencing not only speed but also the very function and nature of their pauses. The detailed procedure presented in the study enables both the replication and the extension of speech sample.

Acknowledgements

This publication was supported by the Cooperatio Program provided by Charles University, research area Linguistics, implemented at the Faculty of Arts of Charles University.

I would like to thank both anonymous reviewers for their valuable comments and recommendations.

REFERENCES

Balkó, I. (2005). K výzkumu tempa řeči a tempa artikulace v různých řečových úlohách. *Bohemistyka,* nr 3, 185–198.

Barik, H. C. (1977). Cross-linguistic study of temporal characteristics of different types of speech materials. *Language and Speech*, *20*(2), 116–126.

Boersma, P., & Weenink, D. (2016). *Praat: doing phonetics by computer* (Version 6.0.23). retrieved 28. 12. 2019 from http://www.praat.org.

Bóna, J. (2014). Temporal characteristics of speech: The effect of age and speech style. *Journal of the Acoustical Society of America*, *136*, Express letters, 116–121.

Bořil, T., & Oceláková, Z. (n.d.). *Počet slabik a AR (rPraat)*. [Number of syllables and AR (rPraat)], [script] retrieved 23. 5. 2024 from https://fonetika.ff.cuni.cz/vyzkum/skripty-a-nastroje.

Dankovičová, J. (2001). *The linguistic basis of articulation rate variation in Czech*. (Forum Phoneticum 71). Hector.

Ferguson, S. H., Morgan, S. D., & Hunter, E. J. (2024). Within-talker and within-session stability of acoustic characteristics of conversational and clear speaking styles. *Journal of the Acoustical Society of America*, *155*(1), 44–55.
Hanžl, V., & Hanžlová, A. (2022). *Prak*. [software] retrieved 2. 5. 2025 from https://github.com/vaclavhanzl/prak.
Hanžl, V., & Hanžlová, A. (2023). Prak: an automatic phonetic alignment tool for Czech. In *Proceedings of XXth International Congress of Phonetics Sciences*. Prague. ID 525, pp. 3121–3125
Huszár, A., & Krepsz, V. (2022). The development of variability in pausing and articulation rate in Hungarian speakers ten years apart. *Govor*, *38*(2), 121–146.
Jacewicz, E., & Fox, R. A. (2010). Between-speaker and within-speaker variation in speech tempo of American English. *Journal of the Acoustical Society of America*, *128*(2), 839–850.
Kohler, K. J. (1986). Parameters of speech rate perception in German words and sentences: duration, F0 movement, and F0 level. *Language and Speech*, *29*, 115–140.
Koopmans-van Beinum, F. J., & van Donzel, M. (1996). Relationship between discourse structure and dynamic speech rate. In *Proceeding of Fourth International Conference on Spoken Language Processing*. https://doi.org/10.1109/ICSLP.1996.607960.
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.
Machač, P., & Skarnitzl, R. (2009). *Fonetická segmentace hlásek*. Nakladatelství Epocha.
Mixdorff, H., Pfitzinger, H. R., & Grauwinkel, K. (2005). Towards objective measures for comparing speaking styles. In *Proceedings of X*th *Speech and Computer – SPECOM 2005*, Patras, Greece, pp. 131–134.
Oceláková, Z. (n.d.). *Počítadlo slabik*. [Syllable calculator] [script] retrieved 23. 5. 2024 from https://keys.shinyapps.io/slabikovac.
Plug, L., Lennon, R., & Smith, R. (2022). Measured and percepived speech tempo: Comparing canonical and surface articulation rates. *Journal of Phonetics*, *95*, 1–15.
Pollák, P., Volín, J., & Skarnitzl, R. (2007). HMM-based phonetic segmentation in Praat environment. In *Proceedings of XIIth Speech and Computer – SPECOM 2007*, pp. 537–541.
Quené, H. (2005). Modelling of between-speaker and within-speaker variation in spontaneous speech tempo. In *Proceedings of Interspeech*, pp. 2457–2460. Lisbon.
Quené, H. (2007). On the just noticeable difference for tempo in speech. *Journal of Phonetics*, *35*, 353–362.
R Core Team (2024). *R: A language and environment for statistical computing* (Version 4.4.1). R Foundation for Statistical Computing. Retrieved 19. 5. 2024 from https://www.rproject.org.
Statistics Kingdom. (2017a). *Correlation Coefficient Calculator* [web application]. retrieved 2. 5. 2025 from https://www.statskingdom.com/correlation-calculator.html.
Statistics Kingdom. (2017b). *Repeated Measures ANOVA Calculator* [web application]. retrieved 2. 5. 2025 from https://www.statskingdom.com/repeated-anova-calculator.html.
Verhoeven, J., De Pauw, G., & Kloots, H. (2004). Speech rate in a pluricentric language: A comparison between Dutch in Belgium and the Netherlands. *Language and Speech*, *47*, 297–308.
Veroňková-Janíková, J. (2005). Dependence of individual speaking rate on speech task. *Acta Universitatis Carolinae – Philologica*, *2005*(1), *Phonetica Pragensia X,* 107–123.
Veroňková, J., & Poukarová, P. (2017). The relation between subjective and objective assessment of speaking rate in Czech radio newsreaders. *Acta Universitatis Carolinae – Philologica, 2017*(3), *Phonetica Pragensia*, 95–107.
Volín, J. (2019). The size of prosodic phrases in native and foreign-accented read-out monologues. *Acta Universitatis Carolinae – Philologica, 2019*(2), *Phonetica Pragensia*, 145–158.
Volín, J. (2022). Variation in speech tempo and its relationship to prosodic boundary occurrence in two speech genres. *Acta Universitatis Carolinae – Philologica, 2022*(1), *Phonetica Pragensia*, 65–81.
Volín, J., Skarnitzl, R., & Pollák, P. (2005). Confronting HMM-based phone labelling with human evaluation of speech production. In *Proceedings of Interspeech 2005*, pp. 1541–1544.
Yuan, J., Liberman, M., & Cieri, C. (2006). Towards an integrated understanding of speaking rate in conversation. In *Proceedings of Interspeech 2006 – ICSLP*, Pittsburgh, pp. 541–544.

RESUMÉ

Příspěvek analyzuje mluvní tempo, artikulační tempo a pauzy ve dvou různých typech žánrů, kterými jsou čtené zprávy a polospontánní projev, v němž se mluvčí představuje přítomnému publiku. Řečový materiál tvoří nahrávky od deseti neprofesionálních mluvčích – žen. Mluvní tempo i artikulační tempo zpráv je vyšší než u představení se, zprávy dále obsahují menší objem pauz, ovšem průměrné trvání pauz je naopak vyšší; všechny uvedené rozdíly jsou statisticky významné. Při rozlišení pauz mezi dílčími zprávami a uvnitř těchto zpráv však rozdíl mezi vnitřními pauzami a pauzami v představení se přestane být významný. Příspěvek dále přináší údaje týkající se variability uvnitř žánrů a mezi mluvčími.

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

2025 ACTA UNIVERSITATIS CAROLINAE PAG. 117–137
PHILOLOGICA 3

INSIGHTS FROM ENGLISH PRONUNCIATION MOOC USERS: THE VIEW FROM 'THE OTHER FORGOTTEN CONTINENT'

ALICE HENDERSON, LAURA RUPP, ADAM WILSON, OLIVIER GLAIN

ABSTRACT
Massive Open and Online Courses (MOOCs) are a valuable contemporary learning resource, and they can be used to promote intelligibility in L2 English pronunciation instruction and learning. Learner reactions to such resources, e.g., in surveys, focus group discussions, comments related to MOOC exercises, etc., potentially reflect their broad language ideology. To this end, we analysed user comments from the MOOC English Pronunciation in a Global World (EPGW) created by Laura Rupp in 2019, focusing on users from Central and South America. This is an under-researched region where three pluricentric languages with different statuses co-exist (Spanish, Portuguese and English), and with a long history of population flows for employment and education. Users' comments from the seven runs of the MOOC reveal how they perceive the notions of fluency and intelligibility, simultaneously providing insights into their aspirations and goals, and thus filling a gap in the research.

Keywords: Central America; fluency; intelligibility; L2 English pronunciation; MOOC; South America

1. Introduction

English is a pluricentric language, like Spanish, Chinese, and several others (see Clyne, 1992), in that it has several standard varieties. It is learned by people of all ages around the world and mastering its pronunciation, while often challenging, is widely accepted as a useful aspect of speaking skills – despite often being neglected in the language classroom[1]. One major difference in learning goals across countries, contexts, and individuals is the degree to which people want to aim for nativelike pronunciation or intelligible pronunciation (see Levis, 2005). In the former, one may want to 'pass' (Piller, 2002), to sound like a native speaker – with all the associated prestige. Aiming for intelligible pronunciation, on the other hand, foregrounds the practical goal of being understandable to others, and often allows for mixes of pronunciation features from the repertoire available to each person involved in an interaction. Intelligibility is therefore more obviously bound by its immediate, ever-changing interactional context, i.e., if we want a person to

[1] See Levis, J. M. (2019). Cinderella no more…

https://doi.org/10.14712/24646830.2025.24

understand us, who do we have in mind? And do we share a language and/or culture with them? What if another person joins our conversation; do we change our speech? This is what Bell refers to as 'audience design', when we shift our speech and pronunciation to adapt to our interlocutors (Bell, 1984).

The choice to aim for nativelike or intelligible pronunciation can be complex for learners as well as for many of the world's 15 million English teachers; according to Freeman et al. (2015), 80% of them are non-native speakers of English. This raises several issues, both conceptual (e.g., What does 'native' mean? Where are the boundaries of an accent?) and practical (e.g., How can teachers help their learners to hear English inside and outside the classroom? Which English? And which resources?). Where possible, teachers and learners choose resources and modes of learning that suit their context and needs, yet they may not always take into account the reality of language diversity beyond the classroom[2]. Learning a pluricentric language forces one to choose, to take a stance in relation to the varieties on offer, whether this means choosing to learn Brazilian Portuguese because one is going to work with people from there (linked to professional goals), or choosing Standard Southern British English because one finds it 'posh' or a favourite aunt had such an accent (reflecting attitudes and emotional preferences).

On-line tools such as Massive Open and Online Courses (MOOCs) are a valuable contemporary learning resource, and they can be used to promote intelligibility in English pronunciation instruction and learning (see, for example, Bueno-Alastuey, 2010). One good example is English Pronunciation in a Global World (EPGW) created by Laura Rupp in 2019 for the FutureLearn platform (Open University, UK). To date, in eight runs of the MOOC, over 134,000 learners from 191 countries and 35 online tutors have participated in the EPGW community, forming an ideal environment for showcasing and experiencing variation in English.

One of EPGW's stated goals is "to explore a variety of different English accents, helping you to understand some of the differences between your pronunciation and that of other English-speaking people" (Rupp, 2019). To this end, participants complete a number of steps associated with various pronunciation activities designed to:

- encourage discussion around notions such as intelligibility and raise awareness of key issues related to pronunciation;
- bring together a large variety of spoken Englishes, to provide maximally varied exposure to English accents and to generate a maximally varied data set;
- provide practice interacting with speakers from around the world, so people learn to handle variation in spoken English.

In this paper we analyse written comments from an exercise at the very start of the MOOC course, where users describe their personal goals for the pronunciation course. Users' comments reveal how they perceive the notion of intelligibility and also reflect their overall language ideology, e.g.: Nowadays it is essential to be understood around the world! Speak English fluently I want to speak like a native, because it can create more opportunities and better interactions with native speakers.

[2] Language diversity amongst learners within a class group will not be touched on in this article, but it can also be seen as a pedagogical resource.

We focus on users from Central and South America for several reasons, partially because it is a region referred to as 'the other forgotten continent' (Friedrich & Berns, 2003). Dedicating a special issue of the journal World Englishes to South America in 2003, the author-editors expressed a hope that:

> … a dialogue with the international research community would empower this region often forgotten and neglected by scientific channels. […] where learning and using English are seen as playing a significant, positive role in the future of the continent. (2003, p. 83)

In line with their hope, and aiming to add to a body of work which has grown only a bit over the past two decades (Friedrich, 2020), we feel it is interesting to understand how L2 English pronunciation is currently perceived in a region where geopolitical realities will continue to evolve, and we sought to use MOOC comments as an entry-point. This paper therefore starts by establishing a theoretical framework (Section 2), before expressing two research questions (§3) and explaining in a methodology section (§4) how we created a corpus and exploited it. In Section 5 we present the results and analysis, followed by a Discussion (§6) and Conclusion.

2. Theoretical framework

2.1 Accessing language ideologies

Language ideology has been defined as "beliefs, or feelings, about languages as used in their social worlds" (Kroskrity, 2004, p. 498). As such, they are "morally and politically loaded representations of the nature, structure, and use of languages in a social world" (Woolard, 2020, p. 2, quoting Irvine, 1989). The author goes on to provide potent examples (2020, p. 2):

> Language ideologies occur not only as mental constructs and in verbalizations but also in embodied practices and dispositions and in material phenomena […] for example, a listener's shudder upon hearing a grating vowel pronunciation, a student's blush at an instructor's attempt to use youth slang, or a speaker's own stammering shame at speaking a language variety she believes she controls imperfectly.

In other words, language ideologies may manifest themselves in our social practices and this manifestation can take place in more or less explicit ways. Another example would be how the decision to use one particular language form rather than another may reflect an ideology, e.g., using a glide instead of a monophthong in the southern US may index an ideology based on rural values. On the other hand, non-participation in the monophtongisation may partake of speakers' identity construction against the rural South (Brunet, 2023).

Language ideologies can sometimes be explicitly thematised in discourse, through what Canut (1998) termed epilinguistic discourse, i.e., stretches of discourse in which representations pertaining to language are co-created, in which these representations are

rendered (quasi)explicit. MOOC comments are a window onto epilinguistic discourse, in which users are (almost) invited by the exercise instructions to share their language ideological stance(s). For example, their comments provide insight into how they conceive of their learning of English, the motivations they have for learning it (especially in a context like the MOOC), etc. Their writing may also reveal how these individuals conceive of, perhaps even define, the (socio)linguistic entity they have embarked on learning, that is the English language e.g., to be intelligible is to be able to speak English like a native speaker. Similarly, comments about wanting to 'avoid a Latino accent' tap into a paradigm of native speakerism.

Compared to other facets of language use, pronunciation is one of the most salient markers of identity we possess and perform, where "accent comes to be used like a badge, showing a person's social identity" (Crystal, 1988, as cited in Mees & Collins 2014, 233). Accents are audible markers of cultural heritage (Hideg et al., 2021) and social interactions can be strongly influenced by them (Gluszek & Dovidio, 2010), for example when they trigger prejudice (Spence et al., 2022). More optimistically, accent in an additional language can also be consciously used to reflect shifts in identity through language learning (Cutler, 2014; Marx, 2002; Piller, 2002).

Given this importance of accent and pronunciation, EPGW potentially contains vast quantities of written comments which can be analysed from a sociolinguistic perspective. Analysing them will reveal how a given language (here, English) is ideologically defined, qualitatively sketching out the perimeters of what is (or is not) acceptable, valuable, desirable when it comes to speaking English[3].

2.2 Choice of geographical region

Central and South America have a long history of north-south contact with the United States (see Casielles-Suárez, 2017; Macías, 2014), constituting a geographic zone rich in English-language educational and employment opportunities, as well as family connections. Flows of people for personal and professional reasons are set to continue, e.g., in 2024, in the top 10 countries of origin for that year's 818,500 naturalized US citizens, Mexico is #1 but the list also includes El Salvador and Colombia[4]. According to the U.S. Department of Labor, in 2024 nearly one-half (48.7%) of the foreign-born labor force was made up of people with Hispanic or Latino ethnicity (Bureau of Labor Statistics, 2024). In general population terms, in 2022 the United States' Hispanic population reached 63.6 million (up from 50.5 million in 2010), making up nearly one-in-five people, up from one-in-twenty in 1970 (Pew Hispanic Center, 2024)[5]. Unfortunately, such statistical realities are accompanied by a long list of negative linguistic and national stereotypes, e.g., "the United States imposes English on Latinos by constructing Spanish speakers as inferior subaltern subjects" (Garcia, 2014, p. 58).

In light of this context, we focus on the comments of MOOC users from Central and South America for three reasons. First, this so-called 'forgotten continent' is large-

[3] See Wilson (2024) for a detailed case study in another context.
[4] See Appendix A.
[5] The Hispanic population is also increasing in Canada, with over a million individuals of 'Hispanic/Spanish-speaking descent' in the 2022 census (3% of the population). See Appendix B for more details.

ly lacking from three widely cited models of Englishes: Kachru's three-circle model of World Englishes (1992, p. 356), Strevens' world map of English (1992, p. 33), and McArthur's circle of World English (1998, p. 97) where only Nicaragua is mentioned. Second, as EFL contexts, they are different to the ESL or EIL settings[6] that tend to get more research coverage, e.g., in Jenkins' (2015) book Global Englishes Central & South American countries are barely mentioned[7].

In general, given that these are all EFL context countries and many of them have a history of trade, immigration, and student flows heading towards the United States, we would expect to find comments linking intelligibility and an American accent – whether as desirable or to be avoided –, and comments about specific professional goals or work. While it is encouraging to see research into teacher beliefs and cognition in this region, (see for example Buss, 2016 about Brazil; Couper, 2016 about Uruguay; Gordon & Barrantes-Elizondo, 2024 about Costa Rica), there is still not much about learners' expectations or hopes with regard to L2 English pronunciation. The current study helps to fill that research gap.

2.3 L2 pronunciation: Fluency and intelligibility ≈ understanding

Intelligibility is a key construct that has been defined in different ways. In line with Derwing and Munro (2015), we define it as what is actually understood. This is typically measured by asking listeners to transcribe what they hear and then counting how many words are correct, though other methods exist (see Kang et al., 2018). Intelligibility is distinct from comprehensibility (i.e., a perception of how difficult it is to understand a speaker) and from accentedness (i.e., a perception of how someone's speech is different to our own or to a type of speech we expect) (Derwing & Munro 2015; also Munro & Derwing, 2020). Non-linguists may use *understandable* synonymously with *intelligible*.

Perhaps one of the most crucial findings in the numerous studies by Derwing, Munro and others over the past 30 years is the fact that one can remain perfectly intelligible even if acoustic features are quite noticeable, i.e., one's accent is quite strong. Given this reality, the aim of EPGW is for learners to develop English pronunciation skills in a world where English is used as a lingua franca. English is used by speakers from a range of different languages, so EPGW focuses on intelligibility rather than nativeness (Levis, 2020) and recognizes personal pronunciation features. Notably, EPGW advocates for listeners to have as much responsibility as speakers for intelligible conversation to occur. The objectives of EPGW are, therefore, for learners to appreciate diversity in English, to speak English that is intelligible to other speakers of English, and to be able to understand other English pronunciations.

Fluency is sometimes confounded with intelligibility, and yet one can improve intelligibility without a perceptible improvement in fluency (Derwing et al., 2014). A basic

[6] EFL refers to English as a Foreign Language, ESL to English as a Second Language, and EIL to English as an International Language.

[7] According to the book's index, Nicaragua & Brazil are only mentioned three times, Mexico twice, and Argentina, Costa Rica, Panama, Surinam only once (Jenkins, 2015). These can be compared to the number of index items referencing English varieties or non-EFL contexts such as: United States (12), American English (9), Australia (6), Australian English (7), India (14), Indian English (9), Nigeria (7), Nigerian English (4).

definition of fluency is "the degree to which speech flows easily without pauses and other dysfluency markers such as false starts" (Derwing & Munro, 2015, p. 177) and it constitutes a positive goal to be attained when learning another language. And while all speakers vary in fluency, non-native speech tends to be less fluent, partly because more time is needed for lexical retrieval (Derwing & Munro, 2015, p. 4). Speech rate is readily perceived by listeners, as people frequently complain that others speak too fast – and sometimes, too slowly. While speaking rate (i.e., the number of syllables produced per second) is indeed one aspect of fluency, hesitation phenomena are another particularly salient aspect. For example, Hilton (2014, p. 34) found that when speaking English, their native French speaking participants hesitated nearly twice as much as the native English speakers, and their fluent runs were shorter. Thus, speed and hesitations are key parts of fluency which individuals notice and comment on, without specialist knowledge.

3. Research questions

A MOOC is conceptualised here as a shared online space where people may write comments and these constitute valid evidence from which to tease out underlying language ideologies. Thus, for EPGW users from Central and South America we have the following research questions:

1. Which themes appear frequently in these MOOC user comments?
2. What do these comments reveal about their underlying language ideology/gies?

Our hypothesis is that the answer to these questions will reflect a regional specificity.

4. Methodology

The course length of the EPGW MOOC is four weeks, with each week having a topic: (1) diversity in English: intelligibility, credibility and identity; (2) English vowel sounds; (3) English consonants, and (4) suprasegmental features in English. Each course week contains a number of learning activities, such as introducing oneself to fellow learners, pronunciation exercises, listening practice, analytical assignments, discussion forums about pronunciation topics, making a recording of your pronunciation and peer-reviewing that of another learner, and reflecting.

For this paper, we focus on a sub-group of total EPGW users (users from countries in Central or South America) and analyse written comments from seven Runs from one exercise (1.8, step 8 in week 1). In Step 1.8, users explicitly describe their personal goals for the pronunciation course and express challenges or concerns regarding their English pronunciation. This step follows ones in which the notions of intelligibility, credibility and identity in English pronunciation have been discussed.

This study builds upon work done by Rupp et al. (2025) which analysed Step 1.8 comments from all EPGW users of Run 1. The four authors used qualitative thematic analysis (e.g., Naeem et al., 2023) to categorize MOOC users' replies to the prompt "formulate concrete pronunciation goals for yourself", identifying one-third of the comments from Step 1.8 as being related to intelligibility. One quarter of such comments from Run 1

came from users in Central or South America: as Run 1 may be atypical, the other Runs needed examining.

The next sub-sections explain how data from the MOOC was extracted to create a corpus, before explaining how the corpus was explored both manually and with text-analysis software.

4.1 Corpus creation

For this study we extracted Step 1.8 comments from all seven Runs (2019-2022) and saved comments only from users who identified themselves as from (or whose IP indicated that they were from) countries in Central or South America; comments containing any explicit mention of being from one of the relevant countries were also included. In this way, we identified 2,169 users from this area of the world, providing 24.6% of all Step 1.8 comments from the seven Runs (Table 1):

Table 1 Number of comments for each run (1–7) produced by EPGW users: All and CAm & SAm.

Run #	1, Feb 2019 (N)	2, Oct 2019 (N)	3, April 2020 (N)	4, Nov 2020 (N)	5, April 2021 (N)	6, April 2022 (N)	7, Oct 2022 (N)	Total # of replies
Total # of participants	11,198	10,260	55,103	20,142	23,837	9,918	6,758	127,982
Total # of comments, Step 1.8	771	593	3,437	1,320	1,617	578	475	8,791[8]
Comments from all except C&S Am	543	427	2562	938	1017	406	309	6,202
Comments by users from C&S Am	207	124	740	336	511	132	119	2,169

We thus conclude that the corpus of comments by users from Central and South America is qualitatively coherent (users all from one geographical zone) and quantitatively substantial (representing one quarter of all comments from Step 1.8). The final corpus has a total of 52,437 words[9].

4.2 Corpus exploitation: Manual and software-aided steps

An initial subset of the 210 comments (30 from each of the seven Runs) was manually coded by the first two authors, before the AntConc software package (Anthony, 2024) was used. The initial manual coding revealed recurrent themes (data-driven), which we then decided to further explore using software (data-informed). Figure 1 visually represents the process.

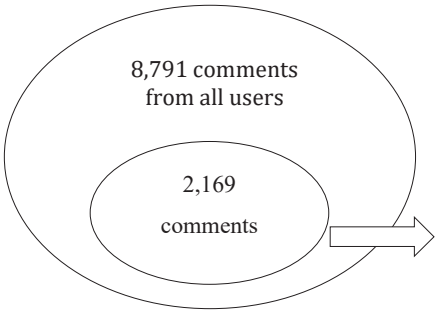
[8] This amount of 8791 is greater than 6202 + 2169 (= 8366) because it included comments from users whose country is 'unknown'.

[9] This number (FileTokens in AntConc) does not indicate how often each word is repeated.

Step 1: Forming First Impressions

Extract comments from regional users

Thematically code & discuss subset of 210 comments



	# of comments to code
Run 1	30 comments
Run 2	30 comments
Run 3	30 comments
Run 4	30 comments
Run 5	30 comments
Run 6	30 comments
Run 7	30 comments

Step 2: In-Depth Exploring with AntConc Software

Generate WordList for each sub-corpus (general overview of occurrence):

– ALLusers_7Runs_comments
– CSAM-only-users_7Runs_comments

Generate KWIC concordances for search terms of interest, e.g., < fluen*>

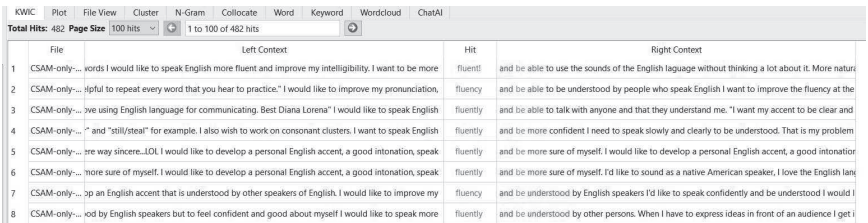


Figure 1 2-step process of exploiting the corpus of comments from Step 1.8.

First, the first two authors manually coded the first 30 comments from each Run, based on the ten categories used by Rupp et al. (2025), such as native speakerism, insecurity, intelligibility, etc. The goal was to get a first impression of the contents and iden-

tify prevalent themes to explore; at this stage, we were open to all frequently occurring themes. After discussing the subset of 210, we decided to focus our further analyses on the theme of intelligibility, because it was the thematically coded category we had attributed most frequently (N = 86/210 for Author 1 and for Author 2, with a few comments coded slightly differently at first[10]). Looking more closely at these comments, we noticed that fluency was explicitly mentioned in many of them, so it became the concrete entry point to the data.

Thus, tools from the software AntConc were used to generate concordance lists (search term <fluen*>[11]) to reveal collocations (lexical associations) in context. The initial results around the notion of fluency showed that it often co-occurs with intelligibility and understanding, so we also generated concordances for those terms (respectively <intelligib*> and <underst*>). Moreover, comments about wanting to be understood or to understand were frequently expressed along with a reason for such a desire, and often linked to a personal or professional goal. Therefore, we generated concordances around types of goals (work, job, study, travel) and for proper nouns, to be able to explore the countries and languages related to these goals. In this method for exploring corpus data – where initial analyses inspire follow-up analyses – each new finding moves the analysis forward toward new terms to examine.

5. Results & analysis

The results will be presented in two parts: first, we discuss comments referring to fluency and understanding (5.1), and then we will focus on those expressing professional aspirations and English-language goals (5.2). In each section we provide some descriptive statistics for the number of occurrences, and look in detail at noteworthy and/or representative comments.

5.1 Fluency & understanding

Our initial observation was that word forms related to the notion of fluency came up 482 times (Appendix C), even though the term is not used in the course before Exercise Step 1.8. Thus, fluency seems to be a tangible concept for these MOOC users.

In the comments from Central and South American users, word forms related to understanding occurred frequently: <understand> (496 occurrences) and <understood> (498 occurrences)[12]. Some expressed a desire to understand others, while some wanted to be understood, with many referring specifically to being understood by more than just native speakers of English, as evidenced by the co-occurrences of understood+by:

[10] Each comment typically was labelled as touching on 2–4 different categories, e.g., intelligibility and native speakerism or credibility.

[11] Characters between < > are used for the exact search terms used in AntConc, with * indicating any character string which follows.

[12] Search Terms: <underst*> and <unde*>. The asterisk makes it possible to find misspelled forms, such as <undestand> (9 occurrences), <undertand> (7 occurrences).

everyone/anyone/other speakers/native and non-native speakers/every English-spoken person:[13]

- *I wish I could be easily understood and not to have a strong accent. My goals are: - being understood by a wide range of English speakers around the world, not just natives speakers;*
- *to develop an English accent that is understood by other speakers of English I would like to improve my accent, so that I could be understood by nativ and non-nativ English speakers;*
- *to improve the pronunciation of some words and also have a greater fluency speaking English so that people understand me. I'd like to be understood by everyone, especially in my career. I would like to improve my pronuntiation to try to sound like a native speaker of the UK.*

Given the frequent reference to these notions (fluency and understanding), we examined whether MOOC users' language highlighted an awareness of intelligibility, and whether they associated being fluent with being intelligible. The proportion of mentions of the term <fluen*> co-occurring with <underst*> and <intelligib*> (Table 2) is different between all EPGW users and the subset of users from Central and South America:

Table 2 Fluency occurrences & co-occurrences: CAm & SAm EPGW users vs ALL other users.

	CAm & SAm Users, 2,169 comments	ALL Other Users, 6,202 comments
Word form	# of mentions	# of mentions
fluen*	482 (22%)	1161 (18.7%)
fluen* + underst* or intelligib*	116 (24%)	137 (11.8%)

MOOC users from this region write about fluency roughly in the same proportion: 22% vs 18.7%. However, they combine <fluen*> with <underst*> or <intelligib*> proportionately twice as much (24%) as all other users (11.8%). This suggests that, for this subset of EPGW users, fluency is linked to being intelligible and understood.

Moreover, fluency is positioned not simply as a goal to be reached, but also as something which is associated with being more confident and even more credible. These ideas co-appear often in the corpus:

- *I would like to be more fluent in order to be more confident. I would like improve my fluency and i would like to develop an English accent that can be understood by other speakers of English;*
- *I would like to improve my English pronuntiation in order to speak more fluently and more confidently, and also be more self-confident about myself at the moment of speaking with foreign;*
- *I would like to improve my pronunciation to be understood and to be confident talking in English I want improve on accent. I want to be more confident when speaking English;*

[13] Comments have been reproduced without any modifications.

- *develop an English accent that is understood by other speakers of English. Pronunciation and fluency, Speak natural, Have credibility. I would like to speak english fluency and my accent can be clearly and understood.*

These comments do not mention native speakers; the goal is to be understood by 'other speakers of English' or with 'foreign'. Other comments explicitly link credibility to a personal accent, e.g., *I would like to be understood and credible in my own accent*. This was, however rare; the expression 'my own accent' occurred 53 times, while 'my personal accent' occurred 20, co-occurring with 'credible' respectively only 7 and 4 times.

The previous examples are typical in that they frame credibility and confidence in the context of being understood: <understood> co-occurs with <credib*> 33 times in a total of 82 comments, and <confid*> 51 times in 231 comments. This is different to explicitly valuing one's ability to understand others, which was quite rare: <to understand> co-occurs only 4 times with <credib*> and 5 times with <confid*>:

- *fluent to sound credible. Also, I would like that my ears get used to the different accents of any country. I would love to be fluent and develop a higher confidence towards recognizing accurate stress in words.*

In this comment the concept of fluency is located in the same stretch of text as both credibility and confidence. However, the key point is that the MOOC user wants their 'ears get used to different accents', i.e., being able to understand others is valued.

5.2 Aspirations and English-language goals

To tap into MOOC users' broader aspirations, plans and even motivation – all topics we hoped they would mention when asked about their English pronunciation goals – we ran two searches (whose results partially overlapped). First, proper nouns of countries and languages in the region were searched (Table 3)[14], given the geographical closeness to North America, the long-established contact with English, as well as the substantial population movements from the south to the north. Then, we also searched for terms related to work, study and travel (Table 4). In both steps, the obvious search candidates were supplemented by items we noticed, as we read through all the comments.

Table 3 presents the results of the proper noun search, to reveal which countries and languages were mentioned in the comments. Abbreviated forms of all the regional countries were searched in AntConc; French and Dutch appeared in a search for capital letters.

In terms of language variety, only two varieties of English are referred to, with <Brit*> being more frequently used than <Amer*> and <US*>:

- *I prefer to maintain my (Dutch) accent, while for my Argentinian students I would prefer to teach them a British or international accent…*

One individual commented on how at school they learned American English but "now I feel the British was lovely" so that had become their goal. Another comment expresses a bit of dilemma between a preference and future employment:

- *I would like to have a British accent, but I am planning to work in the US.*

[14] The search term <English> gave 1314 occurrences, mostly co-occurring with <pronunciation>, so they are not analysed here.

Table 3 Proper nouns used by CAm & SAm EPGW users, in decreasing order of occurrence.

Search term	# of occurrences
Brit*	203
Americ*	152
Span*	45
US*	15
Braz*	12
Portu*	11
Argent*; England	7
Chile	5
Cana*; Mex*	3
French; Urug*; Venez*	2
Dutch	1
Total #	490

There is little mention of local languages (Spanish, Portuguese) and none of other varieties of English. The proper nouns frequently overlapped with aspirations to live and/or work somewhere:

- *… my Brazilian personality. I want to improve my English pronunciation because in my future I would like to live in America, also to join a work where English is the base on speaking. I would like to develop my accent to speak with American people fluently and be understanded.*

Goals were not always expressed in relation to specific countries or language varieties, so as a second step we searched for the terms in Table 4.

Table 4 Comments from CAm & SAm EPGW users with terms related to work, study, and travel.

Search term	# of occurrences
stud*	47
work	44
profession*	38
job*	37
travel	24
live*; opportunit* (22 each)	44
school*; universit* (7 each)	14
clients	4
trainer	2
trip	1
Total #	255

MOOC users were responding to a request to describe their personal goals for the pronunciation course and express challenges or concerns regarding their English pronunciation. The number of occurrences referring to employment (193 = 47 + 44 + 38 + 37 + 22 + 4 + 2) is vastly higher than those referring to travel (25 = 24 <travel> and 1 <trip>) or the 14 occurrences clearly referring to studies, i.e., 7 <school*> and 7 <universit*>. The occurrences of <live> are ambiguous in relation to employment or studies, and thus are not categorised here.

Two comments illustrate how work-related goals may be affective as well as pragmatic, professional:

- *develop an British accent because I lear at school the American and now I feel the British was lovely, it's only personal not for work or thinks like that. I would like to learn those unique features that make pronunciation native-like!*
- *A good English is usefully for my work and this help me to grow up in my personal life.*

Some goals are very precise:

- *I would love to work and care for elderly people in England and I know I need to speak in an understandable way;*
- *to make myself understood … I would like to develop an English accent because I want to work as a reporter at a TV station I would like to speak English in such a way that other people understand me;*
- *I work in a airline company and all the time I need to speak english;*
- *As an actress, I want to learn how can I be understandable.*

Others are more general, about how improving one's pronunciation would be useful on the job:

- *I would like to work in an international environment;*
- *I would like to improve my English performance at work;*
- *I would like to upgrade my English accent because I work with English;*
- *in my new job I have to speak frequently with people that only speak English or French;*
- *I have different needs. First, It is necessary to learn excellently English for my new job. Second, I should be speaking perfect because I will work in New York. Third, I would like to develop an English accent that is understood by other speakers of English.*

Concerning opportunity, almost all of the comments below are examples of people undertaking language investment (Duchêne, 2016), investing time and/or money and/or effort in learning particular language skills in the hope that there will be a return on investment later on, often in the form of new professional opportunities:

- *it can open doors to new opportunities;*
- *I want to improve my pronunciation in English to have better job opportunities;*
- *there are many opportunities in the US, in my field, which is Mathematics;*
- *I want to speak English very well to get better opportunities in my profession, I'm an accountant;*
- *I would like to learn English for the PT test, with this I can have more opportunities to study abroad;*
- *it woul help me to avoid some discrimination problems and would help me to get better opportunities in my job.*

Overall, the comments above frame pronunciation as a key marker of professional success, which is a central motivation in improving one's pronunciation. In general,

this pursuit reflects an instrumental motivation, as described by Gardner and Lambert (1972), where language learning is driven by concrete goals such as professional success or social recognition. Anxiety and insecurity are also expressed in relation to professional contexts:

- *being more and more exposed to foreing clients at my work and I don't feel; confortble yet talking with them. I'm here to work one of my insecurities with my speaking, the pronunciation.*

Such insecurity also appears when teachers' perspectives are explored, because accent and pronunciation are major features of their professional identity. In order to gain credibility, one must approximate a native-like model:

- *to be more confident at my job. I am an English teacher. So, It is important for it.*[15]

That pressure to model nativelike pronunciation – and the insecurity it engenders – is clearly visible in what teachers or tutors wrote:

- *Mu goal is to learn more about how to teach pronunciation to help my students. My personal goal is to break the wall of insecurity when speaking;*
- *I would like to develop an accent that can be attributed to a confident language teacher trainer and trainee;*
- *nowadays english is used as lingua franca (ELF) that's why we have to lead our students into being confortable intelligible when using the foreign language, especially if the aim is communicating with other nonnative speakers.*

The final quote reflects a clear choice to orient teaching by the intelligibility principle (Levis, 2005), rather than nativelike pronunciation.

The desire to work in another country was only voiced four times in the total of 255 comments. Although 22 comments include the verb 'to live', proportionately few (8) mention planning to live in a specific place elsewhere: *I am going to live in the US; before I live in an English-speaking country; to live in London is my biggest dream; live in Switzerland; my dream is to live in an English-speaking country; working towards the opportunity to emigrate to Canada*. Present verb forms are used twelve times, mostly to talk about where oneself lives now (e.g., *I live in a tourist place; the American accent is more noticed where I live*), but two comments are about others (e.g., *my son lives in England; some friends lives there*), and one negative comment is given: *I have never lived in an English-speaking country*. Only two comments refer to the past: *When I lived in the UK; I lived in the USA for five years*.

Living abroad is often a logical extension of studying English at university:

- *I would like to learn English for the PT test, with this I can have more opportunities to study abroad;*
- *to understand the accent of other people in different countries because in the future I want to study and obtein a degree in other country in Europe;*
- *I would like to speak it and understand it very well for my future, because I plan to work when I finish studying, in another country, in a large company.*

[15] See for example Gordon's case study of L2 English pronunciation teachers' identity in Costa Rica (2024).

The verb tenses reveal how individuals assert their agency in looking beyond their current situation to the future, where the United States is not the only destination in their sights:

- *I'd LOVE to have the British accent, I think it's lovely. I'm already studying English at the University, but since one of my biggest dreams is to leave Brazil to live in England, I'll need to improve;*
- *an English accent that can be understood by everyone. I would like to speak more fluently and faster since I am going to study at university to be a translator I need to improve my pronunciation I would like to develop an English accent that is understood by other speakers;*
- *I want to learn how to speak more clearly between english speakers because I want to study an MBA in a contry that the first language it's the english;*
- *My goals are : to be able to enter a university like Harvard and study my specialization or master's degree, travel to Paris, see Niagara Falls, visit the pyramids of Egypt, go to Dubai;*
- *I want to improve my pronunciation so that people can understand me better when I travel or meet foreigners that live in my neighborhood. I think this will give me more confidence.*

Finally, confidence underlies many of their goals, as in the final comment which astutely observes that language is useful not only when traveling, but also in contexts close to home.

6. Discussion

The MOOC EPGW has shown itself to be a suitable environment for enquiring into sociolinguistic issues. Concerning EPGW users from Central and South America, we hypothesized that there would be some regional specificity in the most frequently appearing themes and in the expression of underlying language ideologies.

First, words related to fluency co-occurred with understanding and intelligibility proportionately more frequently among this sub-group of MOOC users, compared to all other users. From a technical, linguistic perspective fluency has no unified scientific definition, and similarly from our analysis of the comments, it does not become clear what EPGW users mean by fluent. Our analysis seems to indicate that they may be equating the nebulous notion of fluency with intelligibility. It is also possible that the EPGW users believe that they are using a word which is in no way fuzzy; it is a very common lay term used to explain language learning goals, to praise someone else's language competency, etc. As shown above, while it is clearly conflated from time to time with intelligibility, fluency also overlaps with the notions of confidence, speed, fluidity, etc. Therefore, we argue that this is another one of those lay terms that functions as a "floating signifier" or "empty signifier", to use a term from critical theory (see Mehlman, 1972; Oxford Reference Overview). This floating quality makes it a quasi-universal goal among language learners, but one that no doubt has different real-world meanings for everyone. We found evidence of MOOC users investing in learning language skills today (language investment) in the hope of more opportunities in future, so their motivation seems to be primarily instrumental.

Second, intriguingly few comments were explicitly linked to the geo-political reality of being an English user in Central & South America. In general, goals were not always associated with specific varieties of English or countries, and the United States was definitely not their primary focus. We had expected there to be far more mentions of wanting to study or work specifically in the United States, given the statistics on mobility related to education and immigration. In reality, references to <Brit*> and <England> (210 = 203 + 7) were slightly more frequent overall, compared to <Americ*> and <US*> (167 = 152 + 15). This may be due to the fact that English more generally is now seen as the global language – more than simply as the language of the USA. Another possibility is that EPGW users know they are taking part in a MOOC which is global and/or explicitly not based in Central/South America. This might orient their responses. Yet another possibility is that some people are not comfortable with disclosing future migration hopes on the internet, especially given American politics around migration from that part of the world – even at the time of the MOOC's Runs.

Third, it may be that in this region of the world, English users are less hampered by a nativist language ideology, as manifested in the many comments of wanting to be understood by more than just native speakers, as well as the absence of comments about purity in other languages or other varieties.

The pedagogical implications are two-fold, one at the institutional level and the other at classroom level. First, Central and South American countries represent EFL contexts in the global English-language teaching landscape. While an intelligibility-focused teaching paradigm has seemingly gained a firm foothold in the published research carried out in ESL contexts in North America or Australia, many EFL contexts remain anchored to the nativeness paradigm. In countries like France, for example, this may be because the competitive exam to become a tenured schoolteacher requires candidates to have native-like pronunciation. The flexible, open-minded goals expressed in the MOOC user's comments encourage us to think that in this region of the world, individuals' perceptions of English have the potential to evolve and absorb the inevitable societal and global changes to come. Second, in teaching contexts a key issue needs to be clearly addressed: who gets to decide whether to focus on achieving intelligibility or nativelike pronunciation. Arguably once people are old enough to put words to their hopes and dreams, open discussion would be useful; learners tend to stay motivated if they have a personal stake in a goal. This holds regardless of whether the context is ESL, EFL or EIL.

In terms of future directions for research, in general, further scientific research on (perceptions of) fluency would be helpful to find out exactly what it is people are referring to – similarly to the notion of intelligibility (Kang et al., forthcoming). More specifically, we would like to explore EPGW users' comments about specific pronunciation features, their beliefs or concerns, and how those evolve over the course of the MOOC, e.g., which specific features are associated with fluency? For example, if a learner mentioned fluency as part of her goals in Step 1.8, in later modules did she change her mind, perhaps become more precise, and explain that in an exercise comment? Similarly, it would be interesting to look at the comments of people who want to improve for professional development and those who need it for travelling, and see how keywords such as *fluency* and *intelligibility* occur, or whether there are any other differences in their comments. It would also be possible to compare our current results with those of other regions (e.g.,

Africa, Asia), to see where nativist language ideology, for example, has a hold. Finally, it would be interesting to look at a subset of MOOC users from all over the world – English teachers and teacher trainers – with regard to their professional identity, because pronunciation plays a central role in this.

7. Conclusion

In 2020 Friedrich updated her sociolinguistic description of the region's "immense diversity – linguistic, ethnic, cultural, musical, geographic, and climatic" (p. 201) in her chapter for the *Handbook of World Englishes*. Her conclusion is bittersweet, in that she (still) finds this landscape underexplored, despite the publication of some works:

> Yet such work, although qualitatively inspiring, remains quantitatively small if compared to the descriptions offered about other areas of the globe, particularly and especially Asia, but also notably Europe. […] There is a great deal of new, creative world Englishes research to be conducted in these fascinating and complex environments. (2020, pp. 201–202)

To conclude, as researchers from outside this area of the world, the analysis of the comments led us to appreciate how much broader are these individuals' views, motivations, and aspirations than our initial preconceptions. Quality and quantity combined to open our eyes to the existing richness and potential of this 'other forgotten continent' and we look forward to further investigations of its sociolinguistic reality.

Acknowledgements

We would like to extend our heartfelt thanks to the MOOC participants, including the devoted mentors. Alice Henderson also received an IRGA grant from UGA, which made it possible for the four authors to meet in person to launch their collaboration. Furthermore, the NRO Comenius Leadership grant awarded to Laura Rupp has made it possible to continue the MOOC, and Silvester Draaijer at VU was unfailingly helpful in applying for – and getting – the grant.

REFERENCES

Anthony, L. (2024). *AntConc* (Version 4.3.1) [Computer Software]. Tokyo, Japan: Waseda University. https://www.laurenceanthony.net/software/AntConc

Bell, A. (1984). Language style as audience design. *Language in Society*, *13*(2), 145–204. https://doi.org/10.1017/S004740450001037X

Brunet, M-P. (2023). "*We have our own idea of vowels in the South*": *A sociophonetic study of /ai/ monophthongization in Middle Tennessee* [Doctoral dissertation, Université Toulouse 2 – Jean Jaurès]. https://theses.fr/2023TOU20098

Bueno-Alastuey, Camino. (2010). Synchronous-voice computer-mediated communication: Effects on pronunciation. *CALICO Journal*, *28*(1), 1–20. https://doi.org/10.11139/cj.28.1.1-20

Bureau of Labor Statistics. (2024). *Foreign-born workers: Labor force characteristics – 2024*. Retrieved from https://www.bls.gov/news.release/pdf/forbrn.pdf
Buss, L. (2016). Beliefs and practices of Brazilian EFL teachers regarding pronunciation. *Language Teaching Research*, *20*(5), 619–637. https://doi.org/10.1177/1362168815574145
Canut, C. (1998). Pour une analyse des productions épilinguistiques'. *Cahiers de praxématique*, 31, 69–90.
Casielles-Suárez, E. (2017). Spanglish: The hybrid voice of Latinos in the United States. *Atlantis*, *39*(2), 147–168. http://www.jstor.org/stable/26426334
Clyne, M. G. (ed.). (1992). Pluricentric languages: Differing norms in different nations. Mouton de Gruyter.
Couper, G. (2016). Teacher cognition of pronunciation teaching amongst English language teachers in Uruguay. *Journal of Second Language Pronunciation*, *2*(1), 29–55.
Cutler, C. (2014). Accentedness, "passing" and crossing. In J. M. Levis & A. Moyer (eds.), *Social dynamics in second language accent* (pp. 145–167). Mouton de Gruyter.
Derwing, T. M., & Munro, M. J. (2015). *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research*. John Benjamins.
Derwing, T. M., Munro, M. J., Foote, J. A., Waugh, E., & Fleming, J. (2014). Opening the window on comprehensible pronunciation after 19 years: A workplace training study. *Language Learning*, *64*(3), 526–548. https://doi.org/10.1111/lang.12053
Duchêne, A. (2016). Language Investment and Political Economy. *Langage et société*, *157*(3), 73–96. https://shs.cairn.info/journal-langage-et-societe-2016-3-page-73?lang=en
Freeman, D., Katz, A., Garcia Gomez, P., & Burns, A. (2015). English-for-teaching: Rethinking teacher language proficiency for the classroom. *ELT Journal*, *69*(2), 129–139. https://doi.org/10.1093/elt/ccu074
Friedrich, P. (2020). South American Englishes and Englishes in South America. In Nelson, C. L., Proshina, Z. G., & Davis, D. R. (eds.), *The Handbook of world Englishes* (2nd ed., pp. 201–214). Wiley Blackwell.
Friedrich, P., & Berns, M. (2003). Introduction: English in South America, the other forgotten continent. *World Englishes*, *22*(2), 83–90. https://doi.org/10.1111/1467-971X.00280
García, O. (2014). U.S. Spanish and education: Global and local intersections. *Review of Research in Education*, 38, 58–80. https://www.jstor.org/stable/43284062
Gardner, R. C., & Lambert. W. E. (1972). *Attitudes and motivation in second language learning*. Newbury House.
Gluszek, A., & Dovidio, J. F. (2010). Speaking with a nonnative accent: Perceptions of bias, communication difficulties, and belonging in the United States. *Journal of Language & Social Psychology*, *29*(2), 224–234. https://doi.org/10.1177/0261927X09359590
Gordon, J., & Barrantes-Elizondo, L. (2024). Idealizing nativeness vs. embracing nonnativeness: A case study on L2 pronunciation teachers' identity. *Language Teaching Research*. https://doi.org/10.1177/13621688241233840
Hideg, I., Shen, W., & Hancock, S. (2021). What is that I hear? An interdisciplinary review and research agenda for non-native accents in the workplace. *Journal of Organizational Behavior*, *43*(2), 214–232. https://doi.org/10.1002/job.2591
Hilton, H. (2014). Oral fluency and spoken proficiency: Considerations for research and testing. In P. Leclerq, A. Edmonds, & H. Hilton (eds.), *Measuring proficiency: Perspectives from SLA* (pp. 28–49). Multilingual Matters.
Jenkins, J. (2015). *Global Englishes: A resource book for students*. Routledge.
Kachru, B. B. (1992). Teaching World Englishes. In B. B. Kachru (ed.), *The other tongue: English across cultures* (2nd ed., pp. 355–366). University of Illinois Press.
Kang, O., Bea, O., & Miao, V. (forthcoming). Timeline on research on L2 speech intelligibility. *Language Teaching*. https://doi.org/10.1017/S0261444825101006
Kang, O., Thomson, R. I., & Moran, M. (2018). Empirical approaches to measuring the intelligibility of different varieties of English in predicting listener comprehension: Measuring intelligibility in varieties of English. *Language Learning*, *68*(1), 115–146. https://doi.org/10.1111/lang.12270
Kroskrity, P. V. (2004). Language ideologies. In A. Duranti (Ed.), *A companion to linguistic anthropology* (pp. 496–517). Blackwell.

Levis, J. M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly*, *39*(3), 369–377. https://doi.org/10.2307/3588485
Levis, J. M. (2019). Cinderella no more: Leaving victimhood behind. *Speak Out! IATEFL PronSIG Journal*, 60, 7–14.
Levis, J. M. (2020). Changes in L2 pronunciation: 25 years of intelligibility, comprehensibility, and accentedness. *Journal of Second Language Pronunciation*, *6*(3), 277–282. https://doi.org/10.1075/jslp.20054.lev
Macías, R. F. (2014). Spanish as the second national language of the United States: Fact, future, fiction, or hope? *Review of Research in Education*, *38*(1), 33–57. https://doi.org/10.3102/0091732x13506544
Marx, N. (2002). Never quite a 'native speaker': Accent and identity in the L2 and the L1, *Canadian Modern Language Review*, *59*(2). https://doi.org/10.3138/cmlr.59.2.264
McArthur, T. (1998). *The English languages*. Cambridge University Press.
Mees, I. M., & Collins, B. (2014). *Practical phonetics and phonology. A resource book for students* (3rd edition). Routledge.
Mehlman, J. (1972). The "Floating Signifier": From Lévi-Strauss to Lacan. *Yale French Studies*, 48, 10–37. https://doi.org/10.2307/2929621
Munro, M. J., & Derwing, T. M. (2020). Foreign accent, comprehensibility and intelligibility, redux. *Journal of Second Language Pronunciation*, *6*(3), 1–27. https://doi.org/10.1075/jslp.20038.mun
Naeem, M., Ozuem, W., Howell, K., & Ranfagni, S. (2023). A step-by-step process of thematic analysis to develop a conceptual model in qualitative research. *International Journal of Qualitative Methods*, *22*, https://doi.org/10.1177/16094069231205789
Oxford University Press: *Floating signifier*. (2025) Oxford Reference. https://www.oxfordreference.com/display/10.1093/oi/authority.20110803095824238
Pew Hispanic Center (Sept 16, 2024). *Key facts about us Latinos*. https://www.pewresearch.org/short-reads/2023/09/22/key-facts-about-us-latinos-for-national-hispanic-heritage-month
Piller, I. (2002). Passing for a native speaker: Identity and success in second language learning. *Journal of Sociolinguistics*, *6*(2), 179–208. https://doi.org/10.1111/1467-9481.00184
Rupp, L. (2019). English pronunciation in a global world [MOOC]. FutureLearn. https://www.futurelearn.com/courses/english-pronunciation
Rupp, L., Simon, C., Henderson, A., Glain, O., & Wilson, A. (2025, October 15–18). A MOOC for pronunciation teaching and research in the real world. [Conference presentation]. PSLLT *16th Pronunciation in Second Language Learning and Teaching Conference.* Concordia University, Montréal, Canada.
Spence, J. L., Hornsey, M. J., Stephenson, E. M., & Imuta, K. (2022). Is your accent right for the job? A meta-analysis on accent bias in hiring decisions. *Personality and Social Psychology Bulletin*, *50*(3). https://doi.org/10.1177/01461672221130595
Strevens, P. (1992). English as an international language: Directions in the 1990s. In B. B. Kachru (ed.), *The other tongue: English across cultures* (2nd ed., pp. 27–47). University of Illinois Press.
Wilson, A. (2024). The ideological construction of the English language in the French agrégation. *E-rea*, 21.2. https://doi.org/10.4000/11wa1
Woolard, K. A. (2020). Language ideology. In J. Stanlaw (Ed.), *The international encyclopedia of linguistic anthropology* (pp. 1–21). Wiley. https://doi.org/10.1002/9781118786093.iela0217

RESUMÉ

Studie se zabývá širší jazykovou ideologií studentů, konkrétně zkoumá jejich postoje ke konceptům *plynulost* a *srozumitelnost*. Metodika výzkumu se opírá o vysoce ceněný internetový zdroj MOOCs (Massive Open and Online Courses) a komentáře jeho uživatelů. Výzkum se soustředí na zdroj zaměřený na osvojování anglické výslovnosti (*English Pronunciation in a Global World* by Laura Rupp) a na uživatele ze zemí Střední a Jižní Ameriky, což je region z tohoto hlediska velmi málo prozkoumaný. Komentáře dávají mimo jiné i nahlédnout na aspirace a cíle studentů a jsou takto využitelné didakticky.

APPENDIX A

Approved Naturalizations for FY 2024 and Top 10 Countries, in thousands.

Country of birth	FY 2024
Mexico	107.7
India	49.7
Philippines	41.2
Dominican Republic	39.9
Cuba	33.7
Vietnam	33.4
China	24.3
El Salvador	21.9
Jamaica	20
Colombia	17.9
All Others	428.8
Total	818.5

Source: USCIS, ELIS. Data accessed October 2024/July 2025.
Note. Due to rounding, the totals may not sum.

APPENDIX B

Immigrant Status and Period of Immigration by Place of Birth (October 26, 2022)

status and period of immigration	Total – Immigrant status and period of immi-gration	Non-immi-grants	Immi-grants	Before 1980	1980 to 1990	1991 to 2000	2001 to 2010	2011 to 20218	2011 to 2015	2016 to 20219	Non-perma-nent residents
Place of birth:											
Central America	239.915	13.185	187.25	10.085	36.665	45.545	41.53	53.42	28.445	24.97	39.48
South America	426.365	10.855	354.395	58.385	49.375	55.715	101.69	89.225	40.85	48.375	61.115

Source: Statistics Canada

APPENDIX C

Occurrences of <fluen*> Mentioned by CAm & SAm EPGW Users

Word form	# of mentions
fluency	181
fluent	126
fluently	161
fluenty	9
fluencitly	2
1x each= fluence, fluencly, fluenlty, fluetly	4
Total #	483

Alice Henderson
Université Grenoble Alpes
alice.henderson@univ-grenoble-alpes.fr

Laura Rupp
Vrije Universiteit Amsterdam
l.m.rupp@vu.nl

Adam Wilson
Université de Lorraine
adam.wilson@univ-lorraine.fr

Olivier Glain
Université Jean Monnet de Saint-Étienne
olivier.glain@univ-st-etienne.fr

2025 ACTA UNIVERSITATIS CAROLINAE PAG. 139–156
PHILOLOGICA 3

VOWEL DURATION IN STRESSED AND UNSTRESSED SYLLABLES IN SPONTANEOUS ENGLISH

NELA BRADÍKOVÁ, RADEK SKARNITZL

ABSTRACT
Many phonetic "truths" are based on descriptions of controlled speech material, and verifying their validity in spontaneous productions is essential. The present study investigates vowel duration as an acoustic correlate of stress in spontaneous English, in communicatively motivated contexts. By analyzing British and American political debates, this study aims to verify previously reported tendencies – specifically, that stressed vowels are significantly longer than unstressed ones. Our analysis of 16 speakers, based on linear mixed effects models, confirms the significant effect of stress on vowel duration and also considers additional factors influencing segmental duration like vowel length, phrase-final position, vowel height, or the nature of the following segment. In addition to stress, multiple regression analysis identifies vowel length, phrase-final position and vowel height as the most influential vowel duration predictors. Despite the variability of spontaneous speech, vowel duration proves to be a reliable correlate of stress, supporting the findings from controlled-speech research.

Keywords: spontaneous speech; lexical stress; vowel duration; English

1. Introduction

Many findings in the speech sciences are based on descriptions of laboratory speech which is more or less controlled: speakers are asked to read sentences or even isolated words or pseudowords, with little linguistic creativity on their part. Such findings have been invaluable for developing theories of speech production, but they tend to be repeated, and it is only rarely that their generalizability is questioned. However, it is conceivable that "language rules" which have been formulated based on more or less tightly controlled speech materials may not hold in spontaneous speech. One of the goals of the present study is therefore to verify some of the claims about the sound patterns of English on speech material which may be regarded as truly spontaneous, naturalistic, and uttered with a clear communicative purpose.

Of course, spontaneous speech constitutes a challenge for researchers at several levels. The phonetic realization of segments may be extremely variable (Greenberg, 1999; Barry & Andreeva, 2001; Nakamura, Iwano, & Furui, 2008). Using Cauldwell's (2013) botanic metaphor, the sound shapes of individual words pronounced in the "jungle" of sponta-

https://doi.org/10.14712/24646830.2025.25

neous speech may very much differ from their canonical forms, which may be observed in the "greenhouse" or "garden". For instance, Johnson (2004) reported that the rate of syllable elision in three- to six-syllabic lexical words ranged between 26 and 59%, or that between 20 and 30% of segments deviated from the canonical form in lexical words longer than four phones (see also a summary of more studies in Tucker & Mukai, 2023). Such levels of reduction may lead to considerable difficulties in performing phonetic alignment at the segmental level (in other words, in identifying individual segments and their boundaries in the stream of speech). In turn, it may then be demanding to extract meaningful data from such material. Nevertheless, we are convinced that the validity of the findings reported in the literature must be put to the test in spontaneous speech. To do so, our present study addresses duration as a correlate of lexical stress in English.

1.1 Lexical stress in English

Correlates of lexical stress represent an area of speech science that has been researched for over 70 years (see van Heuven, 2019, for a summary). Duration is traditionally accepted as the primary acoustic correlate of lexical stress in English. In one of the first studies, Fry (1955) compared the duration and intensity of vowels in noun-verb pairs, such as object or contract in British English, with the target words embedded in sentences. His results showed that both dimensions are important for distinguishing between stressed and unstressed syllables. Lieberman (1960) relied on a similar speech material in American English and examined more acoustic correlates than Fry; he reports fundamental frequency (f_0) and peak amplitude to distinguish between stressed and unstressed syllables, with duration ranking third. The problem is, however, that some of the target words were embedded both in the nuclear and pre-nuclear position in the sentences (e.g., *Kinsey made a survey* and *Let's survey the field*), confounding prominence at the lexical and phrasal levels; the speakers in Lieberman's study were asked to read only the target word, as they would pronounce it in the sentence. In a study on Australian English, Adams and Munro (1978) used read sentences and report duration as "by far the most frequently used cue" (137). Crystal and House (1988) focused specifically on duration in their study of American English; they also relied on read sentences and confirmed duration's role in signalling word stress.

In more recent research, Bettagere (2010) investigated acoustic characteristics of lexical and emphatic stress in American English; he used a word list for the former and simple sentences in which speakers were prompted with a question to place emphatic stress correctly for the latter. Duration was again confirmed to be a more important cue for signalling both levels of stress than *f*0 or amplitude. Fuchs (2016) analyzed acoustic characteristics of lexical and phrasal stress (Fuchs uses the distinction stress and accent, respectively). He used read sentences of Standard British English speakers and found that duration was a correlate of phrase-level but not word-level stress. In what may probably be regarded, from the viewpoint of the analyzed speech material, as one of the most naturalistic studies, Eriksson and Heldner (2015) compared the acoustic characteristics of stressed and unstressed vowels in semi-spontaneous speech, phrase reading, and isolated word reading. Even in this study, however, the semi-spontaneous interview was recorded in a sound-treated studio with the experimenter, without any real-life communicative intent.

The objective of this study will therefore be to determine whether temporal differences between stressed and unstressed syllables – or, more precisely, stressed and unstressed vowels – can be observed in truly spontaneous speech, delivered with a clearly defined audience in the mind of the speakers, namely in a corpus of political debates in British and American English.

1.2 Factors in vowel duration

It is not surprising that the duration of vowels in a language is influenced by multiple factors, and lexical stress is only one of them. That is why this section first summarizes studies concerning various factors that affect vowel duration; many of these have been discussed for instance by Klatt (1976) or van Santen (1992).

Remaining at the suprasegmental level, the duration of segments in general is affected by the position of the word within a prosodic phrase. The most widely examined of these effects is that of phrase-final deceleration (also referred to as phrase-final lengthening), with many studies confirming the finding that syllables at the ends of prosodic phrases tend to be longer in duration (e.g., Lehiste, 1972; Wightman et al., 1992; Byrd & Saltzman, 2003, among others). Specifically, Wightman et al. showed that the lengthening affects the rhyme of the phrase-final syllable and that its degree depends on the depth of the prosodic break; Crystal and House (1988) observed the effect of pauses, but also of final consonants.

Apart from prosodic influences, several factors affecting vowel duration have been documented at the level of individual segments and their interactions. It is logical that vowel length is a crucial factor: phonologically long vowels (including diphthongs) will on average be longer than phonologically short vowels. However, finer distinctions need to be mentioned as well. First, we have to account for what has been called intrinsic vowel duration: open vowels like [a ɑ] are known to have inherently longer duration than close vowels like [I u] (e.g., Peterson & Lehiste, 1960; House, 1961; Solé & Ohala, 2010). Second, vowel duration in English is significantly affected by the phonological voicing of the subsequent consonant. Since English maintains the phonological contrast between voiced and voiceless obstruents in the word-final position (e.g., *meat* and *mead*, *face* and *phase*) and phonetic voicing cannot serve as a reliable cue, the duration of the preceding vowel has become phonologized (Kohler, 1984; Kluender, Diehl, & Wright, 1988). Luce and Charles-Luce (1985) asserted that vowel duration is the most reliable correlate of phonological voicing of word-final stops. While the shorter duration of vowels before phonologically voiceless consonants than before voiced ones seems to be quasi-universal, vowel duration is exploited to a considerably larger extent in English as the major cue to final consonant voicing (Chen, 1970); this phenomenon is typically referred to as pre-fortis shortening.

Returning to our research questions, despite the extensive research on vowel duration and stress in English, there remains a gap in understanding these phenomena in truly spontaneous speech. This study aims to bridge this gap by examining the reported results concerning vowel duration in stressed and unstressed syllables, while also considering multiple segmental and prosodic factors.

2. Method

2.1 Material

As has already been mentioned above, our objective in this study was to analyze spontaneous speech delivered with a real communicative purpose to a clearly defined audience. These requirements are fulfilled by recordings of political debates. At the same time, we aimed to examine standard British and American speech, which would allow us to make at least tentative generalizations about the two major standard variants: Southern British English and the General American accent.

The recordings were obtained from publicly accessible archives of the BBC programme Westminster Hour (www.bbc.co.uk/programmes/b006s624) and C-SPAN network (www.c-span.org) for the British and American recordings, respectively. The material consisted of recordings of 16 speakers in total: eight British English speakers (four females, four males) and eight American English speakers (four females, four males). Within each group, neither the British nor the American speakers displayed any significant regional features in their speech.

For each speaker, the speech material analyzed in this study spanned approximately 200 words of spontaneous speech, which amounts to roughly 60–100 seconds per recording. In total, the material provided 4,927 tokens of stressed and unstressed vowels. The speech material is summarized in Table 1, which shows the number of words and vowels depending on the number of syllables.

Table 1 Summary of the analyzed speech material.

	number of syllables	word tokens	vowel instances
monosyllabic words	1	2507	
polysyllabic words	2	632	1264
	3	260	780
	4	74	296
	5	16	80

2.2 Analysis

The recordings were first transcribed and automatically segmented using the P2FA forced-alignment tool (Yuan & Liberman, 2008). Overlapping speech was excluded from subsequent analyses. All speech sound boundaries were manually adjusted based on phonetically motivated criteria (Machač & Skarnitzl, 2009). In the next stage, all syllables had to be labelled as either stressed or unstressed. Naturally, we were interested in actual realizations of stress, and not in canonical or dictionary forms; in some treatments, one would say that the task was to identify accented syllables (but see a different use of the term accent in Section 1.1). To do this, we used careful auditory analysis, relying on an alternation of broader- and narrower-context listening. Since the material consisted of spontaneous speech, it was natural that some ambiguous cases appeared, and these were resolved in a joint analysis session by both authors. Lastly, we identified prosod-

ic boundaries – in ToBI (Beckman & Ayers Elam, 1997), these would correspond to both major (BI4) and minor (BI3) prosodic phrases – to distinguish between syllables in the phrase-final and non-final stress groups. All these analyses were conducted in Praat (Boersma & Weenink, 2024), which was also used to extract vowel durations and other relevant information using a dedicated script.

To evaluate the statistical significance of stress and other segmental and prosodic factors on vowel duration, we built a linear mixed-effects (LME) model using R (R Core Team, 2024) and the *lme4* package (Bates et al., 2015). Initially, we constructed a model with absolute vowel duration as the dependent variable which, however, when checked for homoscedasticity, suggested that the data were heteroscedastic. Of course, that is not surprising, since duration values are known to be positively skewed. Therefore, we subsequently constructed a model using log-transformed duration values as the dependent variable. As fixed effects, we included STRESS (with the levels being stressed or unstressed), the quality of the following segment (coded as FOLLOWING, with the levels open syllable, voiced coda, voiceless coda), VOWEL LENGTH (short, long, diphthong), VOWEL HEIGHT (close, mid, open), phrase-FINAL position (phrase-final or -internal), WORD LENGTH, and the VARIETY of English (British, American). Finally, as random effects, we included by-SPEAKER and by-WORD intercepts, as well as by-SPEAKER random slopes for the effect of STRESS. The complete structure of the random effects was thus (1 + Stress | Speaker) + (1 | Word); this accounts for the possibility that every speaker treats the difference between vowel durations in stressed and unstressed syllables differently. The statistical significance of the fixed effects was ascertained using likelihood ratio tests, comparing the fit of the complete model to that of a model without the given predictor. When appropriate, post-hoc pairwise comparisons were conducted using the *emmeans* package (Lenth, 2024): estimated marginal means (emmeans) were computed from the LME model, with Tukey's method applied to adjust for multiple comparisons.

A linear mixed effects model and likelihood ratio tests inform us about the significance of the individual factors – that is, about whether the factors' contribution to the overall model is significant. However, we were also interested in the relative contribution of the individual predictors, in how they compare in determining the final duration of the vowels. For that reason, we conducted a series of stepwise multiple linear regression (MLR) analyses using the *lm* function in R. Log-transformed vowel duration was used as the dependent variable, and the same factors as with LME served as predictors. We employed backward elimination and stepwise selection in both directions, using the Akaike Information Criterion (AIC) as the guiding metric to evaluate the contribution of each predictor to the model.

All plots were generated using the *ggplot2* package (Wickham, 2016). Note that the results of LME modelling are based on the log-transformed values of vowel duration. However, for the sake of more transparent interpretability of the results, the boxplots used to illustrate the effect of individual variables will depict absolute vowel durations.

3. Results and discussion

3.1 Statistical modelling of effects on vowel duration

The main objective of this study was to ascertain the difference in the duration of vowels depending on whether they appear in stressed or unstressed syllables in spontaneous speech. Our findings are consistent with previous research (see Section 1), with STRESS found to have a highly significant effect on vowel duration ($\chi^2(1) = 47.2$, $p < 0.0001$); detailed results of LME modelling are provided in the Appendix. Stressed vowels are indeed longer in duration than unstressed vowels, across both varieties of English, as can be seen in Figure 1. Note that the figure shows absolute duration values, whereas the statistical model is based on the log-transformed values of duration (see Section 2.2).

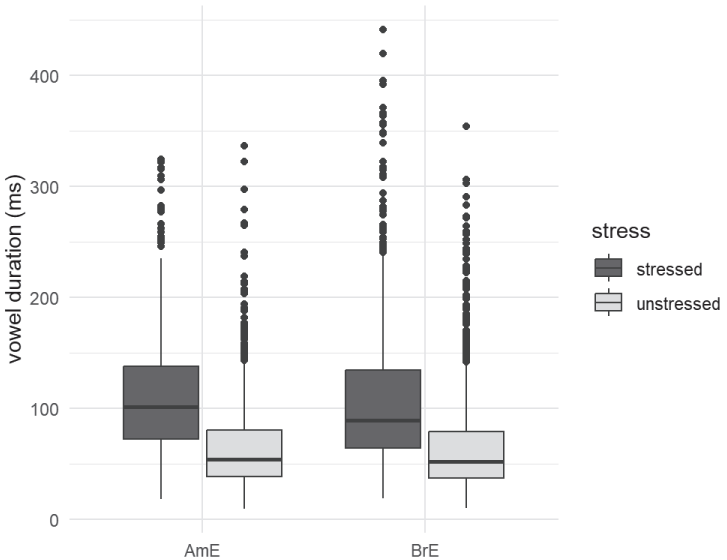


Figure 1 Absolute duration of stressed and unstressed vowels in the two varieties.

Vowel durations in stressed and unstressed syllables were not a priori expected to differ between British and American English. Although, based on inspecting Figure 1, there seems to be a small difference in the duration of stressed vowels, with those in American English slightly longer than in British English, VARIETY did not have a significant effect on vowel duration in our model ($\chi^2(1) = 2.7$, $p = 0.1$). Speaker identity was treated as a random factor in the analysis, but as seen in Figure 2, there were no prominent differences between the productions of individual speakers: the absolute duration of stressed vowels was consistently longer than that of unstressed vowels across speakers, and to a rather similar degree. It is interesting to observe the range of outlier values seen in Figure 2; this further showcases the variability and complexity of spontaneous speech. Given that no effect of VARIETY was observed in our model, this factor will not be included in subsequent visualizations.

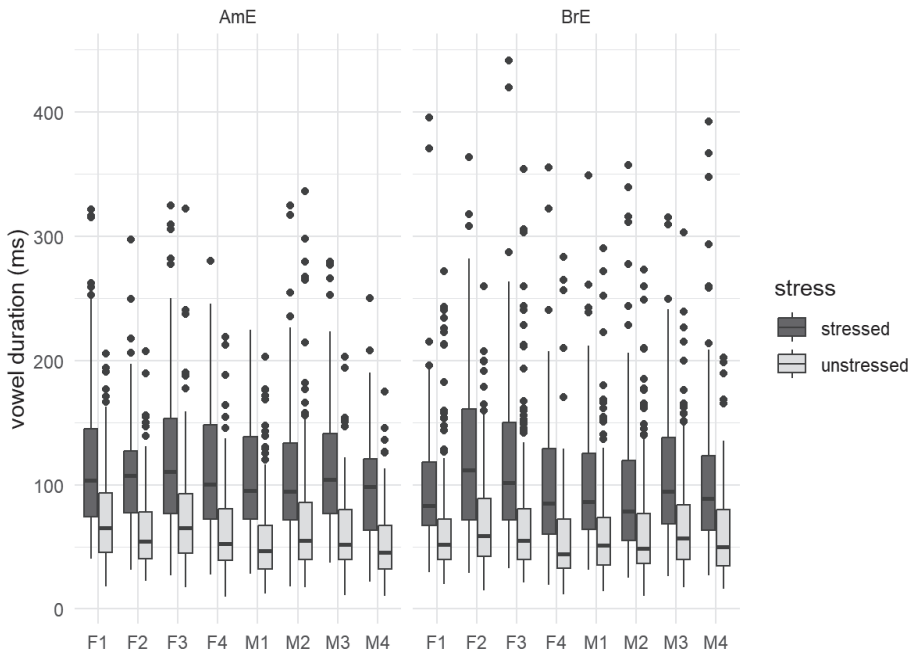


Figure 2 Absolute duration of stressed and unstressed vowels in individual speakers of the two varieties.

Along with stress, proximity to a prosodic boundary is one of the suprasegmental factors which is known to influence vowel duration (see Section 1.2). In our analysis of spontaneous speech, phrase-FINAL position proved to have a significant effect on (log-transformed) vowel duration ($\chi^2(1) = 661.4$, $p < 0.0001$). The effect can also be observed in the absolute duration values shown in Figure 3. As the graph shows, vowels are longer in the phrase-final position, and the difference is even more evident in stressed vowels; this is supported by the significance of the interaction between phrase-FINAL position and STRESS ($\chi^2(1) = 65.9$, $p < 0.0001$). It is important to note that we did not distinguish between prosodic break types (BI3 and BI4) or between phrase-final vowels followed by a pause and those occurring within a longer stretch of speech. It can be assumed that the presence of a pause could further increase the difference between the duration of phrase-final and phrase-internal vowels (both stressed and unstressed). This assumption would however be worth confirming through further research.

The next factor considered in our analysis was VOWEL LENGTH. Originally, we operated with a two-level factor (short and long vowels, where the latter included diphthongs). However, as shown in Figure 4, diphthongs turned out to be considerably longer than long monophthongs in our data. For that reason, a three-way distinction was used in the model, which proved that VOWEL LENGTH has a significant effect on vowel duration ($\chi^2(2) = 573.0$, $p < 0.0001$). Figure 4 also shows that the effect of phonological vowel length is consistent across stressed and unstressed vowels, and the interaction between STRESS and VOWEL LENGTH is not significant ($\chi^2(2) = 2.0$, $p > 0.3$).

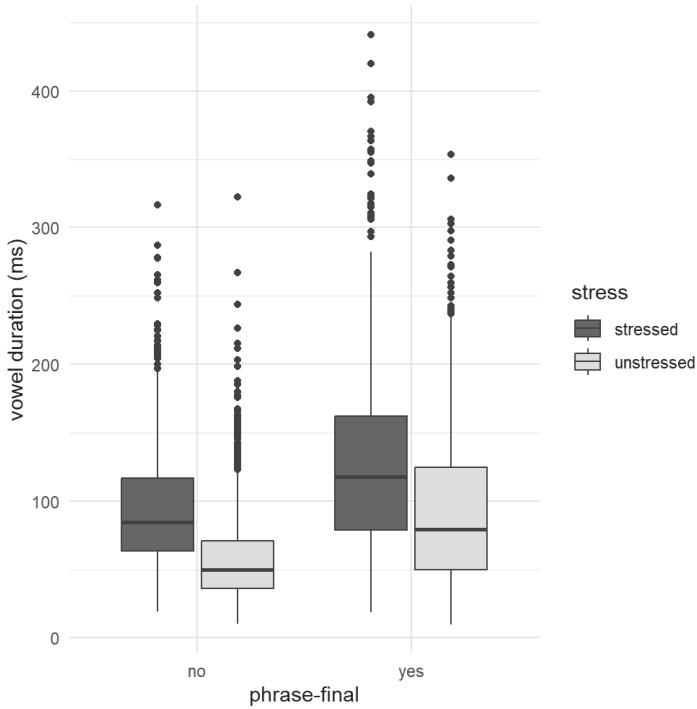


Figure 3 Absolute duration of stressed and unstressed vowels in phrase-final and phrase-internal positions.

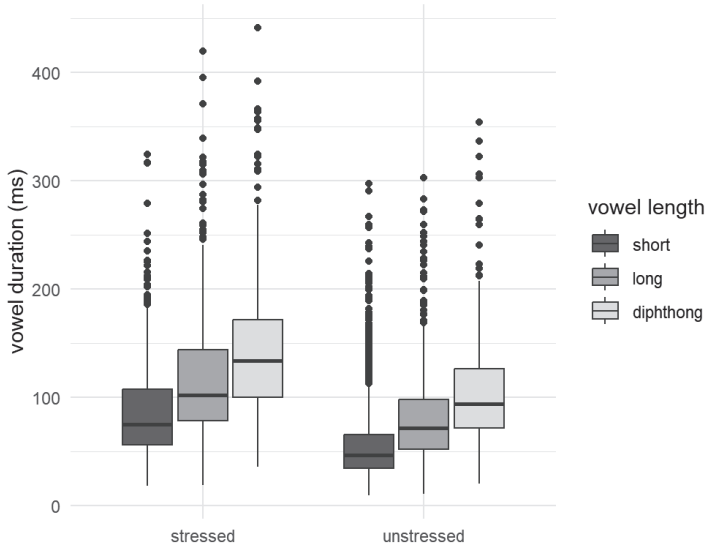


Figure 4 Absolute duration of stressed and unstressed vowels, depending on vowel length.

Another effect that we considered was intrinsic vowel duration, coded as VOWEL HEIGHT in this analysis, which was found to be significant ($\chi^2(2) = 184.4$, $p < 0.0001$). The results shown in Figure 5 suggest that the relationship between vowel duration and vowel height differs within stressed and unstressed syllables. This is also confirmed by the significant interaction between STRESS and VOWEL HEIGHT ($\chi^2(2) = 40.4$, $p < 0.0001$). To be specific, vowel duration varies systematically across the three height categories in stressed syllables in accordance with tendencies reported in the literature (see Section 1.2). However, within unstressed syllables, mid vowels are shorter in duration than high vowels, although a post-hoc pairwise comparison, calculated using the *emmeans* function with Tukey's adjustment, shows that the difference falls short of statistical significance ($p > 0.2$). This tendency can presumably be explained by the fact that the unstressed mid vowels include a lot of schwas, which are likely to be most reduced not only spectrally but also in the temporal domain, and thus would be shorter in duration than high vowels.

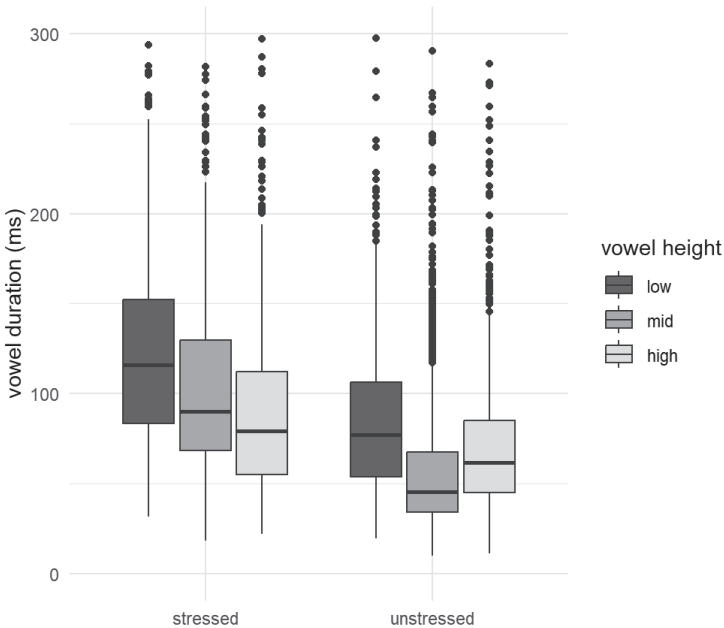


Figure 5 Absolute duration of stressed and unstressed vowels, depending on vowel height.

The next effect we were interested in examining was that of WORD LENGTH, expressed as the number of syllables in a word. Since English is a language traditionally referred to as stress-based, which involves the temporal compression of unstressed syllables within a stress group, increasing word length should, other things being equal, be reflected in shorter vowel durations. Our analysis supports this: the effect of WORD LENGTH turned out to be significant ($\chi^2(1) = 75.8$, $p < 0.0001$). More detailed results, presented in Figure 6, reveal a clear difference between the absolute durations of vowels in monosyllabic and polysyllabic words. More specifically, it is particularly stressed vowels in monosyllabic words which are longer in duration than those in polysyllabic words. However, post-hoc

pairwise comparisons show that both stressed and unstressed vowels in disyllabic words are significantly longer than those in four-syllabic words ($p < 0.05$); the comparison of di- and three-syllabic words was only marginally significant ($p = 0.08$).

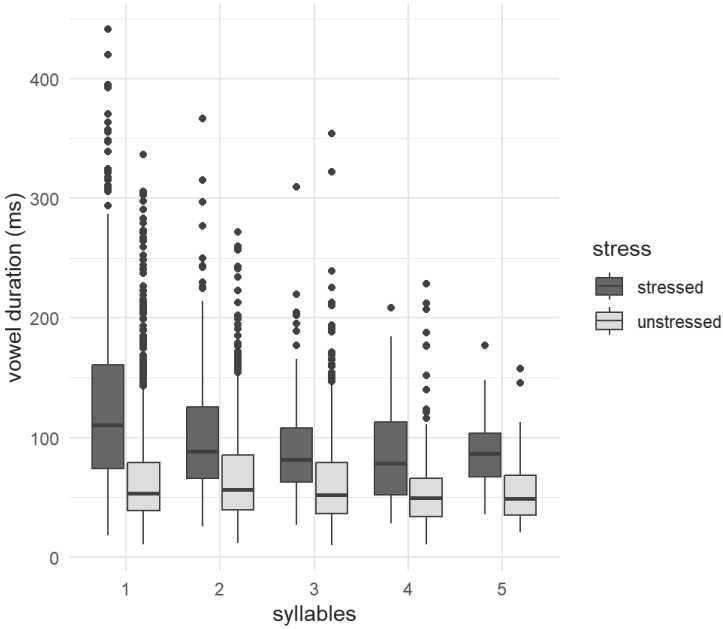


Figure 6 Absolute duration of stressed and unstressed vowels, depending on the number of syllables in a word.

The last factor, whose effect on vowel duration we aimed to examine, was the nature of the FOLLOWING segment. As described in Section 1.2, English uses vowel duration to cue the difference between fortis and lenis obstruents in the coda. It is therefore not surprising that our spontaneous speech data confirm the significance of this factor ($\chi^2(2) = 52.7$, $p < 0.0001$). Note that the voiced group comprises both lenis obstruents and sonorants in the coda (i.e., *maid* as well as *main*). A more detailed analysis of Figure 7 suggests that the significance of the factor FOLLOWING is mostly due to the longer duration of vowels in open syllables (e.g., *May*), rather than by pre-fortis shortening; however, post-hoc pairwise comparisons (conducted on the log-transformed values of the LME model) confirm the statistical significance between pre-voiced and pre-voiceless (fortis) vowels in both stressed and unstressed syllables as well ($p < 0.001$).

It is interesting to probe the duration of pre-fortis and pre-voiced vowels further because, despite the significance of the pairwise comparisons reported above, we expected to see a greater difference in a contrast which cues phonological distinctions with a high functional load. In Figure 8, we only show the absolute duration data for monosyllabic words, and it is obvious that the duration difference is considerably more pronounced in the stressed syllables. In other words, the effect of pre-fortis shortening is most salient in stressed monosyllabic words (*mate* ['mĕɪt] and *maid* ['meɪd̥]), while it may

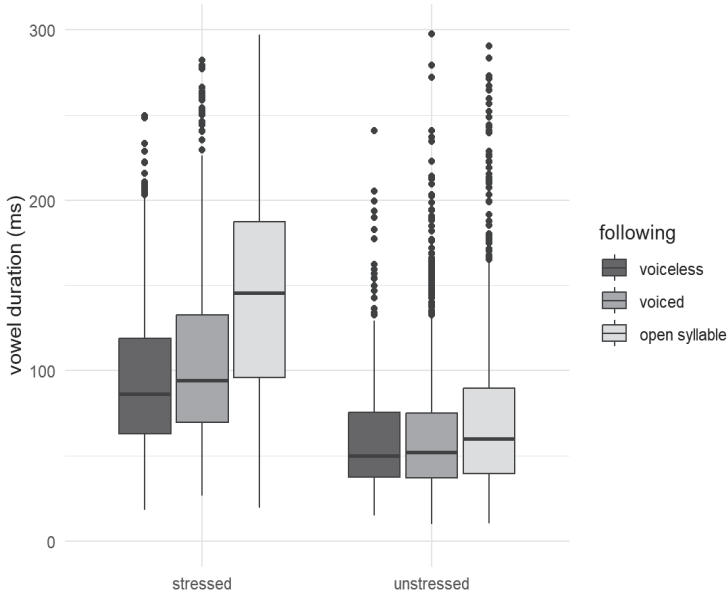


Figure 7 Absolute duration of stressed and unstressed vowels, depending on the nature of the following segment.

be less salient in longer word pairs (for example, *fickle* ['fĭkɬ] and *figure* ['fɪgə] or *sightline* ['săɪtlaɪn] and *sideline* ['saɪdlaɪn]).

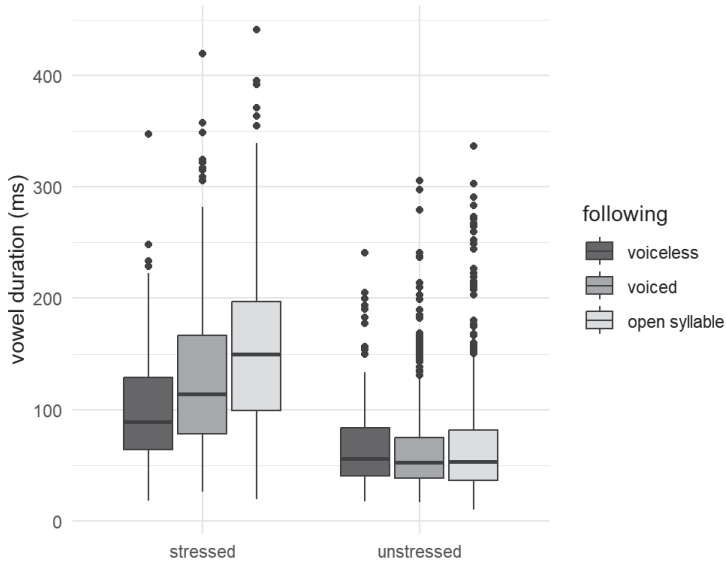


Figure 8 Absolute duration of stressed and unstressed vowels in monosyllabic words, depending on the nature of the following segment.

Having discussed the significance of each of the factors individually using LME modelling, in the next section we will assess their relative importance for vowel duration using multiple regression analysis (MRA). The drawback of MRA is, however, that random factors like the speakers' identity are not considered.

3.2 Relative contribution of predictors to vowel duration

As mentioned in Section 2.2, the log-transformed vowel duration functioned as the dependent variable in the MR analysis. We used the Akaike Information Criterion (AIC) to assess the relative importance of the factors affecting vowel duration. The AIC compares different models (i.e., models with different sets of predictors) to each other, whereby the model with the lowest AIC is regarded to be the best trade-off between accuracy (goodness of fit) and complexity (the number of predictors). An accompanying metric, frequently reported in MLR analysis, is R^2_{adj} (the adjusted coefficient of determination), which is used to evaluate the explanatory power of a regression model; more specifically, it corresponds to the percentage of variance explained by the model, adjusted for the number of predictors used in the model.

The overall model, with all predictors included, was significant: $F(10, 4909) = 395.4$, $p < 0.0001$, $R^2_{adj} = 0.445$. In other words, the complete model explained 44.5% of the variance in the vowel duration data. Through a bidirectional stepwise analysis, the metric ranks the predictors according to the degree of explained variance in the MLR model from highest to lowest. The AIC values of predictors and the corresponding R^2_{adj} values, listed in the order of explained variance, are presented in Table 2 and visualized in Figure 9.

Table 2 Results of bidirectional stepwise MLR, with AIC and R^2_{adj} values in decreasing order of explained variance (see text).

predictor	AIC	R^2_{adj}
intercept	−5145.9	
vowel length	−6430.4	0.230
phrase-final	−7178.6	0.339
stress	−7724.5	0.408
vowel height	−7906.9	0.430
syllables	−8005.5	0.441
variety	−8019.0	0.443
quality of the following segment	−8032.5	0.445

Starting with the null model with only the intercept value, adding vowel length results in a considerable decrease in AIC, indicating that vowel length alone accounts for a significant amount of explained variance of vowel duration, approximately 23%. The addition of the phrase-final condition further lowers the AIC and increases the degree of explained variance to 33.9%. Incorporating the affiliation of the vowel to a stressed vs. unstressed syllable adds another nearly 7% of explained variance, and subsequently including vowel height brings the explained variance to 43%.

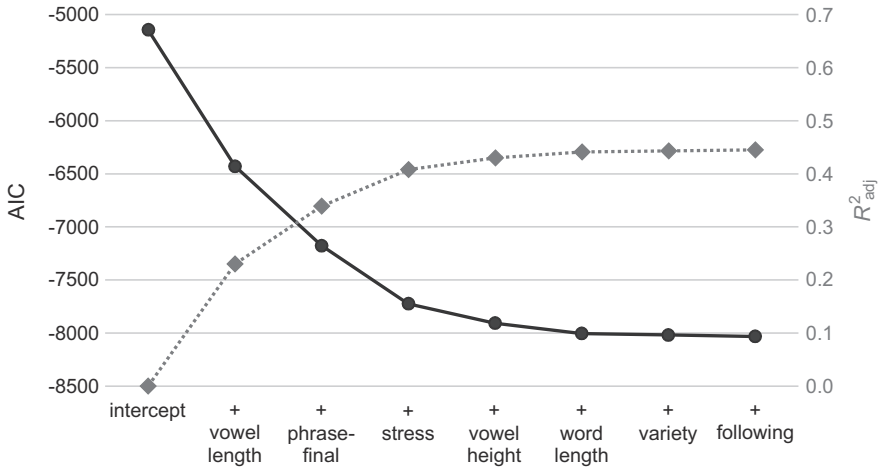


Figure 9 Changes in the Akaike Information Criterion (AIC, in black circles, axis on the left) and the adjusted coefficient of determination (R^2_{adj}, in grey diamonds, axis on the right) in the multiple regression model (see text).

The remaining three predictors only provide minor changes to AIC and therefore contribute less new information to the model. If we were aiming for a parsimonious but still effective model of vowel duration in our spontaneous English data, the following four predictors should be included:

- vowel length: the distinction between short monophthongs, long monophthongs, and diphthongs affects vowel duration to the largest extent
- position within the phrase: vowels in the last stress group of a prosodic phrase are significantly longer than phrase-internal vowels
- stress: vowels in stressed syllables are longer than those in unstressed syllables
- vowel height: the correlate of intrinsic vowel duration, with open (low) vowels longer than close (high) vowels

4. General discussion

The present study examined vowel duration as the primary acoustic correlate of lexical stress in English. Its primary aim was to determine whether the long-held relationship between stress and vowel duration, which was observed on more or less controlled speech materials, would be verified in truly spontaneous speech, namely in political debates broadcast in the United Kingdom and the United States. In the most general sense, the study thus examined whether stressed vowels are longer in duration than unstressed vowels even in spontaneous English. In addition, we wanted to see how other factors, known to affect the duration of vowels, modulated their relationship with lexical stress.

Our results confirm that vowel duration may be considered an important cue for the distinction between stressed and unstressed syllables; this relationship holds both glob-

ally (Figure 1) and at the level of individual speakers (Figure 2). At the same time, no difference was observed between our British and American speakers, although Figure 1 suggests a tendency for stressed syllables to be shorter in British English; this may confirm the impression that the British politicians were, overall, speaking slightly faster than the American ones.

As for the modulating factors on the segmental level, our analysis included the effect of vowel length (considering the temporal differences of short and long monophthongs and diphthongs), intrinsic vowel duration (conceptualized as vowel height), and the shortening of vowels before fortis consonants (which is used in English to cue the voicing contrast in the syllabic coda). All three segmental factors were found to affect vowel duration, in line with the findings reported in the literature (see Section 1.2). However, the relationship between these factors and stress contrast differs. The duration difference between stressed *and* unstressed vowels holds for vowel length (Figure 4), and the importance of this factor is confirmed by its first ranking in the stepwise multiple regression analysis (cf. Table 2 and Figure 9). With vowel height, the exceptionally short nature of *schwa* vowels blurs the relationship somewhat but, overall, the relationship also holds (Figure 5). The effect of coda fortis consonants was the least obvious: it is manifested to the greatest extent in the stressed syllables of monosyllabic words (Figures 7 and 8).

Regarding prosodic factors that modulate the effect of lexical stress on vowel duration, we focused on phrase-final deceleration and word length. Our results suggest that vowels in phrase-final stress groups were longer than those in phrase-internal stress groups, and this difference affects both stressed and unstressed vowels (Figure 3). In comparison, vowel duration operates slightly differently in monosyllabic words, with particularly stressed vowels being considerably longer in monosyllabic words (see Figure 6).

While this study examined the relations between the presence of a prosodic boundary and vowel duration, it is important to note that we did not differentiate various depths of prosodic boundary (ToBI 3 and 4) or between phrase-final stress groups that were followed by a pause and those that were immediately followed by another prosodic phrase. For a complete understanding of phrase-final deceleration, it would be worth adding another level of the phrase-FINAL factor, which would correspond to vowels in prepausal stress groups, and determining whether its effect on vowel duration in spontaneous speech would be greater.

As emphasized throughout this study, the purpose of our research was to analyze real spontaneous speech. Since spontaneous speech is a highly complex phenomenon and many differences have been observed between canonical forms and spontaneous realizations, we had expected our main findings to be less clear-cut. However, our results demonstrate that duration is indeed a crucial cue to the stressed vs. unstressed distinction in English spontaneous speech – naturally, along with other correlates like peripheral vs. reduced (centralized) vowel quality.

Unsurprisingly, the analysis of spontaneous speech proved to be methodologically laborious; this concerns both phonetic segmentation (where segments one expects may not be realized at all) and the identification of stressed (accented) syllables. It is obvious that a meticulous auditory and acoustic analysis (see Section 2) was essential to carry out these steps and make our analysis valid. Although the extent of the speech material which could be examined for this study was relatively restricted, we believe that it was sufficient

for the results to be reliable and that the presented study provides interesting insight into stress and vowel duration in spontaneous English.

REFERENCES

Adams, C., & Munro, R. R. (1978). In search of the acoustic correlates of stress: Fundamental frequency, amplitude, and duration in the connected utterance of some native and non-native speakers of English. *Phonetica*, *35*, 125–126. https://doi.org/10.1159/000259926

Barry, W., & Andreeva, B. (2001). Cross-language similarities and differences in spontaneous speech patterns. *Journal of the International Phonetic Association*, *31*(1), 51–66. https://doi.org/10.1017/S0025100301001050

Bates, D., Mächler, M., & Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beckman, M.E., & Ayers Elam, G. (1997). *Guidelines for ToBI labelling*, version 3. The Ohio State University Research Foundation.

Bettagere, R. (2010). Differences in acoustic characteristics of stress patterns in American English. *Perceptual and Motor Skills*, *110*(2), 339–347. https://doi.org/10.2466/pms.110.2.339-347

Boersma, P., & Weenink, D. (2024). *Praat: Doing phonetics by computer (Version 6.4)*. Retrieved from www.praat.org

Byrd, D., & Saltzman, E. (2003). The elastic phrase: modelling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics*, *31*, 149–180. https://doi.org/10.1016/S0095-4470(02)00085-2

Cauldwell, R. (2013). *Phonology for listening: Teaching the stream of speech*. Speech in Action.

Chen, M. (1970). Vowel length variation as a function of the voicing of the consonant environment. *Phonetica*, *22*, 129–159. https://doi.org/10.1159/000259312

Crystal, T. H., & House, A. S. (1988). Segmental durations in connected-speech signals: Syllabic stress. *Journal of the Acoustical Society of America*, *83*, 1574–1585. https://doi.org/10.1121/1.395912

Eriksson, A., & Heldner, M. (2015). The acoustics of word stress in English as a function of stress level and speaking style. *Proceedings of Interspeech 2015*, 41–45. https://doi.org/10.21437/Interspeech.2015-9

Fry, D. B. (1955). Duration and intensity as physical correlates of linguistic stress. *Journal of the Acoustical Society of America*, *27*(4), 765–768. https://doi.org/10.1121/1.1908022

Fuchs, R. (2016). The acoustic correlates of stress and accent in English content and function words. *Proceedings of Speech Prosody 2016*, 435–439. https://doi.org/10.21437/SpeechProsody.2016-89

Greenberg, S. (1999). Speaking in shorthand – A syllable-centric perspective for understanding pronunciation variation. *Speech Communication*, *29*, 159–176. https://doi.org/10.1016/S0167-6393(99)00050-3

House, A. S. (1961). On vowel duration in English. *Journal of the Acoustical Society of America*, *33*(9), 1174–1178. https://doi.org/10.1121/1.1908941

Johnson, K. (2004). Massive reduction in conversational American English. In K. Yoneyama & K. Maekawa (eds.), *Proceedings of the first session of the 10th international symposium on spontaneous speech: Data and analysis* (pp. 29–54). The National International Institute for Japanese Language.

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, *59*, 1208–1221. https://doi.org/10.1121/1.380986

Kluender, R. K., Diehl, R. L., & Wright, B. A. (1988). Vowel length differences before voiced and voiceless consonants: An auditory explanation. *Journal of Phonetics*, *16*, 153–169. https://doi.org/10.1016/S0095-4470(19)30480-2

Kohler, K. J. (1984). Phonetic explanation in phonology: the feature fortis/lenis. *Phonetica*, *41*, 150–174. https://doi.org/10.1159/000261721

Lehiste, I. (1972). The timing of utterances and linguistic boundaries. *Journal of the Acoustical Society of America, 51*(6B), 2018–2024. https://doi.org/10.1121/1.1913062

Lenth, R. (2024). *emmeans: Estimated Marginal Means, aka least-squares means*. R package version 1.10.3. Retrieved from https://CRAN.R-project.org/package=emmeans
Lieberman, P. (1960). Some acoustic correlates of word stress in American English. *Journal of the Acoustical Society of America*, *32*, 451–454. https://doi.org/10.1121/1.1908095
Luce, P. A., & Charles-Luce, J. (1985). Contextual effects on vowel duration, closure duration, and the consonant/vowel ratio in speech production. *Journal of the Acoustical Society of America*, *78*(6), 1949–1957. https://doi.org/10.1121/1.392651
Machač, P., & Skarnitzl, R. (2009). *Principles of phonetic segmentation*. Epocha.
Nakamura, M., Iwano, K., & Furui, S. (2008). Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance. *Computer Speech and Language*, *22*, 171–184. https://doi.org/10.1016/j.csl.2007.07.003
Peterson, G. E., & Lehiste, I. (1960). Duration of syllable nuclei in English. *Journal of the Acoustical Society of America*, *32*, 693–703. https://doi.org/10.1121/1.1908183
R Core Team (2024). *R: A language and environment for statistical computing* (version 4.4.2). R Foundation for Statistical Computing. Retrieved from https://www.r-project.org
Solé, M. J., & Ohala, J. J. (2010). What is and what is not under the control of the speaker: Intrinsic vowel duration. In C. Fougeron, B. Kühnert, M. D'Imperio & N. Vallée (eds.), *Laboratory phonology 10* (pp. 607–655). De Gruyter Mouton. https://doi.org/10.1515/9783110224917.5.607
Tucker, B. V., & Mukai, Y. (2023). *Spontaneous speech*. Cambridge University Press.
van Heuven, V. J. (2019). Acoustic correlates and perceptual cues of word and sentence stress: towards a cross-linguistic perspective. In R. Goedemans, J. Heinz, & H. van der Hulst (eds.), *The study of word stress and accent: theories, methods and data* (pp. 15–59). Cambridge University Press. https://doi.org/10.1017/9781316683101.002
van Santen, J. P. H. (1992). Contextual effects on vowel duration. *Speech Communication*, *11*, 513–546. https://doi.org/10.1016/0167-6393(92)90027-5
Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag.
Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. J. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America*, *91*(3), 1707–1717. https://doi.org/10.1121/1.402450
Yuan, J., & Liberman, M. (2008). Speaker identification on the SCOTUS corpus. *Proceedings of Acoustics '08*, 5687–5690.

APPENDIX

Results of the linear mixed-effects (LME) model

```
Linear mixed model fit by REML ['lmerMod']
Formula: logdur ~ stress + variety + following + vowel_length + vowel_heigh
t + final + word_length + (1 + stress | speaker) + (1 | word)
   Data: data

REML criterion at convergence: 5592.1

Scaled residuals:
     Min  1Q       Median   3Q       Max
 −4.2128  −0.6209  0.0008   0.6171   4.0259

Random effects:
 Groups      Name               Variance    Std.Dev   Corr
                                            .
 word        (Intercept)        0.028624    0.16919
 speaker     (Intercept)        0.007411    0.08609
             stressunstressed   0.003076    0.05546   0.10
 Residual                       0.159165    0.39895
Number of obs: 4920, groups:  word, 1271; speaker, 16

Fixed effects:
                     Estimate   Std. Error  t value
 (Intercept)          5.011059    0.047249  106.055
 stressunstressed    −0.304312    0.020069  −15.163
 varietyBrE          −0.083741    0.047242   -1.773
 followingvoiced     −0.140363    0.023037   −6.093
 followingvoiceless  −0.180285    0.024891   −7.243
 vowellengthlong     −0.220653    0.026139   −8.441
 vowellengthshort    −0.513550    0.022568  −22.756
 vowelheightlow       0.212663    0.023517    9.043
 vowelheightmid      −0.065959    0.017435   −3.783
 finalyes             0.399319    0.014973   26.670
 wordlength          −0.075482    0.008593   −8.784
```


RESUMÉ

Řada fonetických „pravidel“ je založena na analýzách kontrolovaného materiálu nebo z dat získaných cíleně pro experimentální účely. Pro úplné porozumění těmto přijímaným „pravdám“ je však nezbytné ověřit jejich platnost i ve spontánní řeči. Tato studie se zabývá trváním samohlásek jako akustickým korelátem slovního přízvuku v angličtině ve spontánní řeči produkované v komunikačně motivovaných kontextech. Za cíl si klade ověřit dřívější poznatky o souvztažnosti přízvuku a trvání vokálů, a to v nahrávkách osmi amerických a osmi britských mluvčích účastnících se politických debat. Konkrétně studie ověřuje hypotézu, že přízvučné vokály mají signifikantně delší trvání než vokály nepřízvučné. Pomocí lineárních smíšených modelů naše analýza potvrzuje statisticky významný vliv slovního přízvuku na trvání vokálů. Výzkum zároveň zohledňuje další faktory ovlivňující segmentální trvání, jako jsou

fonologická délka daných vokálů (tedy zda se jedná o krátký či dlouhý monoftong nebo diftong), pozice v rámci prozodické fráze, vokalická výška jakožto korelát inherentního trvání nebo charakter následujícího segmentu. Z krokové vícenásobné regresní analýzy vyplývá, že kromě přízvuku jsou nejvýznamnějšími prediktory trvání vokálů fonologická délka vokálu, finální pozice v prozodické frázi a vokalická výška; tyto čtyři faktory vysvětlují 43 % variability v trvání našich samohlásek. Navzdory vysoké variabilitě charakteristické pro spontánní řeč se trvání vokálů ukazuje jako spolehlivý korelát přízvuku, což potvrzuje dřívější závěry získané z výzkumu kontrolované řeči.

Nela Bradíková
Institute of Phonetics
Faculty of Arts, Charles University
Prague, Czech Republic
bradikovanela@gmail.com

Radek Skarnitzl
Department of Czech Language and Literature
Faculty of Education
University of Hradec Králové
radek.skarnitzl@uhk.cz

2025 ACTA UNIVERSITATIS CAROLINAE PAG. 157–175
PHILOLOGICA 3

ACOUSTIC ANALYSIS OF VOWELS IN CZECH DISYLLABIC WORDS PRODUCED BY L1-GERMAN SPEAKERS

ANNA CHABROVÁ

ABSTRACT
This study presents the results of an acoustic analysis of vowels produced by L1-German speakers learning Czech as a foreign language and provides a brief overview of vowel behaviour in Czech and German. The analysed vowels are /a aː ɪ iː u uː/ in Czech disyllabic words, and the speakers are eight women with varying levels of proficiency. Vocalic formants F1 and F2 were analysed, and the differences in formant values between long and short vowels were calculated. Furthermore, vowel duration was measured, and differences between the durations of long and short vowels were assessed. The results are compared with reference values for female native speakers of Czech and German and indicate that speakers in the present study do not sufficiently distinguish between short and long vowels in Czech. Additionally, the absolute vowel durations observed in this study are longer than the reference values for Czech speakers.

Keywords: Czech as L2; German as L1; vowel length; vowel duration; vowel formants; disyllabic words; acoustic analysis

1. Introduction

Pronunciation is one of the key areas that speakers must master when learning a new language. Incorrect pronunciation can significantly hinder communication and may even lead to complete misunderstanding. The present study is part of a broader research project conducted at the Institute of Phonetics at the Faculty of Arts, Charles University, which focuses on the phonetic characteristics of non-native Czech speech. Experimental findings have confirmed, among other things, difficulties in the realization of vowel length in non-native speakers whose first language is Russian, Ukrainian, or Polish (e.g., Palková et al., 2020; Veroňková & Bořil, 2020a, 2020b; Veroňková et al., 2020). The present experiment focuses on vowel length in Czech disyllabic words produced by native German speakers.

Vowels in German and Czech differ in several aspects. Czech has five short monophthongs, five long monophthongs, and three diphthongs. Short and long vowels form phonologically distinctive pairs, and all Czech vowels are primarily lax and non-nasalized (Palková, 1994: 172). The short vowels are /ɪ e a o u/ and the long vowels are /iː eː aː oː uː/. Vowel pairs /e eː/, /a aː/ and /o oː/ do not differ significantly in quality, but there are

https://doi.org/10.14712/24646830.2025.26

noticeable differences in the vowels /ɪ iː/ and /u uː/. In the case of /ɪ iː/, the difference is so substantial (cf. e.g. Podlipský et al., 2009; Skarnitzl & Volín, 2012; Paillereau & Chládková, 2019) that separate transcription symbols are now used to distinguish the short and long variants not only for comparative purposes, but also for transcription within the Czech phonological system. Skarnitzl and Volín (2012) worked with recordings of several dozen university students reading a text in a recording studio, and their results serve as reference values for Czech native speakers. Therefore, I use their study as a reference in this paper.

Regarding vowel quantity, older publications report that long vowels are twice as long as short ones (Palková, 1994: 179), but more recent findings suggest a different picture. Skarnitzl (2012), who analysed recordings of professional speakers in broadcast news and worked with non-normalized durations in his study, determined the following duration ratios between the long and short variants of each vowel (see Table 1, on the left). Paillereau and Chládková (2019) who analysed normalized durations in spontaneous speech produced by non-professional speakers, reported different ratios, particularly for /ɪ iː/ (see Table 1, on the right). Their participants came from diverse social backgrounds and the recordings were based on spontaneous speech.

However, both studies agree that long vowels are less than twice as long as their short counterparts, and that the durational contrast is smaller for /ɪ iː/ and /u uː/ compared to the other vowel pairs. The reason is the qualitative difference between the short and long variants of these vowels, which facilitates perceptual discrimination and thus allows for a reduction in durational contrast (Skarnitzl, 2012: 151).

Table 1 Duration ratios between the long and short vowel in each vowel pair according to Skarnitzl (2012) and Paillereau and Chládková (2019).

	V: / V ratio	
vowel pair	Skarnitzl	Paillereau & Chládková
iː / ɪ	1.29	1.66
eː / e	1.72	1.78
aː / a	1.79	1.73
oː / o	1.73	1.87
uː / u	1.60	1.65

In German, the situation is different and vowel quality and quantity are closely interconnected. Unlike in Czech, vowel length in German is not indicated by diacritical marks. German also has more vowel qualities than Czech, and these are often directly linked to a specific realization of length, e.g., the vowel /ʊ/ is always phonologically short (Becker, 2012: 31). However, some vowel qualities exist in both short and long forms, and in such cases the speaker must know whether the vowel in a given word should be pronounced long or short, e.g., the vowel /u/ can occur in both long and short forms (Becker, 2012: 31).

Vowel quality and quantity in German are also linked to word stress. The distribution of certain vowels is restricted to either stressed or unstressed syllables, e.g., the tense long vowel /i:/ occurs only in stressed syllables, whereas the reduced vowel /ə/occurs only in unstressed ones (Kleiner & Knöbl, 2015: 32). According to many authors (e.g., Kleiner & Knöbl, 2015: 32), only short vowels can occur in unstressed syllables in native German words. However, vowel duration in German should be considered not only in absolute (i.e., short vs. long) but also in relative terms. Some authors (e.g., Becker, 2012; Jessen, 1993) assume that stressed syllables may be relatively longer than unstressed ones, and that this relatively longer duration contributes to their prominence. This may have significant implications for L1-German speakers learning Czech, as such a situation does not occur in Czech.

In Czech, word stress is fixed, typically falling on the first syllable of the word, and it has no influence on vowel length or quality (Ashby & Maidment, 2015: 135; Skarnitzl, 2008: 199–200). Both short and long vowels can occur in stressed as well as unstressed syllables (e.g., /pla:nɪ/ 'plans' and /plani:/ 'wild'). In the case of disyllabic words, German and Czech typically agree in placing stress on the first syllable (Kleiner & Knöbl, 2015: 59).

The present study is part of a broader experiment, the aim of which was to obtain data on the realization of vowel length and quality in Czech disyllabic words produced by native German speakers, using both a perception test and acoustic analysis. Deviations from the canonical form are expected, and different structural types are likely to exhibit different patterns of behaviour.

This paper presents the results of the acoustic analysis of vowel duration and formant measurements. The results of the perception test, along with summarised findings on vowel duration, were already published in Chabrová & Veroňková (2022). The present study expands and completes the analysis of vowel duration.

The study by Chabrová & Veroňková (2022) confirmed the assumption that vowel length in Czech disyllabic words poses a challenge for native German speakers – native Czech speakers perceived only 42% of the words in accordance with the original text, which was read by the German speakers. For the remaining 58%, Czech listeners perceived length other than the originally correct and intended one. A summary of the perception test results is presented in Table 2. Rows, labelled original, indicate vowel length in the original disyllabic word as written in the text read by German speakers. Columns, labelled perceived as, show the vowel length perceived by Czech listeners in the perception test. Bolded cells correspond to words perceived in agreement with the original text, i.e., pronounced correctly.

Table 2 Overall percentage distribution of how the original structures SS/SL/LS/LL were perceived. S = short vowel, L = long vowel. In Chabrová & Veroňková (2022).

(%)	perceived as SS	perceived as SL	perceived as LS	perceived as LL
original SS	**40.3**	25.1	21.8	12.8
original SL	28.1	**27.1**	26.7	18.1
original LS	12.3	9.0	**64.8**	13.9
original LL	17.5	14.1	32.5	**35.9**

Speakers achieved the highest success rate with the LS structure (65%), i.e., a long vowel in the first, stressed syllable and a short vowel in the second, unstressed syllable. On the contrary, the lowest success rate was found for the SL structure (27%), where the first syllable contained a short vowel and the second syllable a long vowel. The success rate for short vowels alone was 65% and for long vowels alone, 61%. The following factors influenced the success of vowel realization: the type of short/long structure in the target word (i.e., short-short = SS, short-long = SL, long-short = LS, long-long = LL), vowel quality, the position of the vowel in the first or second syllable (for the analysed items, this data could only be obtained for vowels /ɪ iː/, since in the material used, /a aː/ occurred only in the first syllable and /u uː/ only in the second syllable). The frequency of occurrence was mapped for the word forms used and their lemmas. Due to the smaller size of the spoken language corpus, which did not include many of the target items, the SYN2020 corpus of written language was used (Křen et al., 2020). The frequency of the lexemes/word forms may have influenced the evaluation in individual cases, but did not affect the overall trends.

2. Method

2.1 Speakers

The source material consisted of read-aloud recordings from eight native speakers of German. All participants were women[1] (aged 21–38) with varying levels of Czech language proficiency (five intermediate speakers, estimated level A2–B1 according to the CEFR; three advanced speakers, estimated level B2–C1 according to the CEFR). The intermediate speakers had been learning Czech for one to two years at the time of recording, had completed at least one year of Czech Studies, and were residing in the Czech Republic during the recording period. The advanced speakers had been learning Czech for four or more years at the time of recording and were long-term residents of the Czech Republic. One of the speakers was from Austria, and the remaining seven were from Germany.

2.2 Material

The recordings[2] were made in the studio of the Institute of Phonetics at the Faculty of Arts, Charles University, using an AKG C 4500 B-BC microphone. The audio was record-

[1] Only female speakers were included in the experiment, as recordings from too few male speakers were available, and their voices would have been too easily identifiable among the female speakers in the perception test.

[2] The recordings used in this study came from two sources: original recordings made by the author and recordings from the corpus of non-native Czech speech compiled at the Institute of Phonetics, Faculty of Arts, Charles University. The material for this corpus was recorded as part of the Czech science foundation grant project GA ČR 18-18300S *Zvukové vlastnosti češtiny v komunikaci nerodilých a rodilých mluvčích*. The recordings conducted by the author served as an extension of the corpus using identical texts.

ed at a sampling rate of 48 kHz with 16-bit quantization and was saved and processed in WAV format.

The recording text consisted of sentence pairs. Each pair included a first sentence, which served merely to establish a context, and a second sentence containing the target phenomenon. The two sentences within each pair were semantically related, whereas different sentence pairs were not related to each other. Each pair was presented on a new line, and the speakers read one A4 page at a time, meaning they could only see one sheet at once. They were allowed to read and prepare this sheet just before recording, then handed it in and received the next one for preparation. This procedure was chosen to minimize the risk of speakers noticing patterns between words or sentences and identifying the focus of the study. At the same time, the short preparation ensured fluent delivery with minimal hesitations or errors. Informal interviews conducted after the recordings confirmed that the speakers did not identify the focus of the study.

Suitable carrier sentences containing the target words were selected from the recordings. The target items were disyllabic words forming groups of four, three or two that differed only in vowel length (e.g., a group *sazi*, *sazí*, *sází*). All syllables in the target words were open, to avoid any potential influence of syllable structure (open vs. closed) on vowel duration. A total of 27 carrier sentences were used, each containing one target word in an unambiguous context. The target words were never located at the very beginning or end of a sentence, and sentence pairs containing target words from the same group (e.g., the group *sazi*, *sazí*, *sází*) were not placed close to one another to mask the target phenomenon.

The set of target words includes the following items:

a) 3 groups of four: myli, milí, mýlí, mílí; platu, platů, plátu, plátů; valy, valí, vály, válí;
b) 3 groups of three: sazi, sazí, sází; spali, spály, spálí; vazu, vazů, vázu;
c) 3 groups of two: kraji, krájí; planý, plány; sliby, slíbí.

An example of a group of four words platu / platů / plátu / plátů used in carrier sentences:

(1) Nebyla na tom špatně. K jeho **platu** dostala ještě podporu od pojišťovny. (SS)
*She wasn't doing badly. In addition to his **salary**, she also got support from the insurance company.*
(2) Ředitel se dohodl s odbory. V prosinci dostali šest **platů** jako bonus. (SL)
*The director reached an agreement with the unions. In December, they received six **salaries** as a bonus.*
(3) Desku tvaroval podle vzoru. Na jednom **plátu** oceli pracoval dva dny. (LS)
*He shaped the plate according to the pattern. He worked on one steel **plate** for two days.*
(4) Potřebovali vyplnit mezeru. Dohromady spojili pět **plátů** železa. (LL)
*They needed to fill the gap. Altogether, they joined five iron **plates**.*

Despite their different spelling, the letters y/ý and i/í are pronounced identically in Czech: y and i are both realized as the vowel [ɪ], while ý and í are both pronounced as [iː]. Similarly, the letters ú and ů both represent the long vowel [uː]; the difference between them is only graphical and depends on the vowel's position within the word.

Target words were extracted from the recordings, and after excluding items with disturbing, irremovable noise, a set of 203 stimuli was obtained, which was an adequate

number for the length of the perception test. The set contains a balanced representation of the four possible combinations of vowel length in disyllabic words (SS, SL, LS, LL) from all eight speakers. The vowel quality combinations in the target words are a–i, a–u, and i–i. Words containing e and o were intentionally excluded from the experiment because the long vowels /o:/ and /e:/ are located on the periphery of the phonological system (Vachek, 1968: 30–34).

2.3 Acoustic analysis

For the acoustic analysis, vowel durations in the words from the perception test were measured using Praat (Boersma & Weenink, 2020), and their formant values were obtained through automatic extraction followed by manual verification. For duration measurements, vowel boundaries were marked in two different ways. The first segmentation approach followed the recommended guidelines for segmenting phonemes (Machač & Skarnitzl, 2009), with vowel boundaries placed according to the formant structure. The second segmentation approach placed greater emphasis on perception and was guided visually by oscillogram: in relevant cases, the vowel boundary was shifted to include a voicing offset following the end of the vowel's articulation itself (Machač & Skarnitzl, 2009: 136–137).

Vowel duration was normalized relative to the articulation rate of the entire word using the method employed by Veroňková & Bořil (2020b), in order to allow for comparisons across different speakers. For each target word, the articulation rate in syllables per second was calculated, based on which the average articulation rate for each speaker was determined. Normalized duration was obtained by multiplying the actual duration of the word by average articulation rate of the respective speaker, and the resulting value was then divided by the overall average articulation rate across all speakers.

In the acoustic analysis of formants, only the first variant of segmentation was used, that is, the one according to Machač & Skarnitzl (2009). Formants F1 and F2 were first measured automatically in Praat in the middle third of vowel duration. The formants were extracted using the Burg method (time steps: automatic; maximum number of formants: 5; formant ceiling: 5500 Hz; window length: 0.025 s; pre-emphasis from: 50 Hz) using a script (Bořil, 2015). Since errors can occur in the automatic extraction of formant values, the obtained values were manually checked by comparing them with reference data. The reference used was the study by Skarnitzl & Volín (2012), which includes data from dozens of speakers and offers high-quality reference values for male speakers, including standard deviations. Based on these values, I calculated reference ranges for F1 and F2 of the respective vowel for female speakers. Any values outside these ranges were then manually verified and, if necessary, corrected. The reference ranges were calculated as follows: one standard deviation was added to and subtracted from the average male formant values, resulting in a reference range for male speakers. In accordance with generally observed patterns, formant frequencies for female speakers are approximately 15–20% higher than those for male speakers (Skarnitzl & Volín, 2012: 8). In order to obtain the reference ranges for female speakers, I increased the values of the male reference ranges by 17.5%.

3. Results

3.1 Formant analysis results

Table 3 presents the results of formant analysis in both hertz and ERB. For values in ERB, the percentage difference between formants of short and long vowels is also given. The conversion to ERB was performed using formula[3] $21.4 \times \log10(0.00437 \times f + 1)$ (Glasberg & Moore, 1990). Columns labelled F1/F2 difference show the difference in F1/F2 between short and long variants of the respective vowel.

Table 3 Mean F1 and F2 values in the analysed material.

	F1 (Hz)	F2 (Hz)	F1 (ERB)	F2 (ERB)	F1 difference (% of ERB)	F2 difference (% of ERB)
a	840.27	1457.23	14.33	18.56	2.16	0.84
a:	877.22	1429.26	14.64	18.41		
ɪ	370.81	2540.66	8.95	23.17	5.12	0.70
i:	343.30	2589.92	8.52	23.34		
u	373.64	1024.67	9.00	15.81	5.67	5.88
u:	343.15	911.70	8.51	14.93		

Differences between the formants of short and long variants of the respective vowels are small. For /a a:/, the long variant has a slightly higher F1, while there is almost no difference in F2. In case of /ɪ i:/, the short variant has a somewhat higher F1 and a lower F2, but in absolute terms, these differences remain small. The most differentiated are vowels /u u:/, where the short variant has both higher F1 and F2, and is therefore slightly more centralized than the long variant.

Tables 4a and 4b present formants of vowels /ɪ i:/, divided according to whether the vowel occurs in the first or the second syllable.

Table 4a Mean values of F1 and F2 for vowels /ɪ i:/ in the *first* syllable.

1st syll.	F1 (Hz)	F2 (Hz)	F1 (ERB)	F2 (ERB)	F1 difference (% of ERB)	F2 difference (% of ERB)
ɪ	362.05	2531.45	8.82	23.14	3.41	0.80
i:	343.87	2587.26	8.53	23.33		

[3] For the conversion from Hz to ERB, an online converter was used (ERB-rate scale converter, n.d., University College London. Retrieved 5. 6. 2025 from http://www.homepages.ucl.ac.uk/~sslyjjt/speech/erb.html).

Table 4b Mean values of F1 and F2 for vowels /ɪ iː/ in the *second* syllable.

2nd syll.	F1 (Hz)	F2 (Hz)	F1 (ERB)	F2 (ERB)	F1 difference (% of ERB)	F2 difference (% of ERB)
ɪ	373.42	2543.39	8.99	23.18	5.63	0.67
iː	343.13	2590.72	8.51	23.34		

While values for long /iː/ are the same in the first and second syllable, the values for short /ɪ/ differ slightly. In the second syllable, the short vowel shows slightly higher F1 and F2 values, resulting in a slightly greater percentage difference between the short and long vowel in F1 compared to the first syllable. Overall, however, the differences between the first and second syllable remain small.

The formant values of all vowels are graphically displayed in Figure 1.

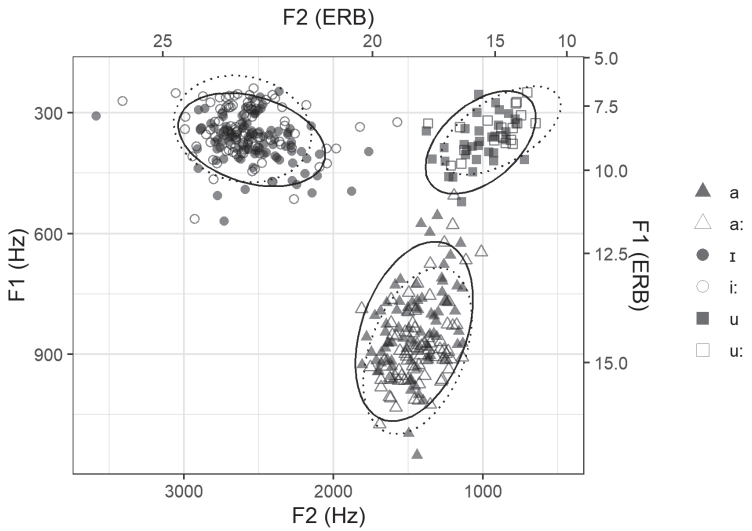


Figure 1 Formant values of vowels in Hz and ERB; ellipses cover 95% of the values. Solid ellipses represent short vowels and dotted ellipses represent long vowels.

The results were further filtered in Tables 5a–5d by combinations of vowel quality and quantity in order to analyse whether the vowel quantity patterns SS (short-short), SL (short-long), LS (long-short), and LL (long-long) differ from each other. Values are presented in both hertz and ERB.

F2 formants for /a aː/ are again nearly identical, but differences can be observed in F1. While in the SS pattern the values are closer to those of the short variant, in SL, LS, and LL patterns they correspond more to the long variant. This would be expected in the LS and LL patterns, where a long vowel is indeed supposed to occur, but in the SL pattern, this represents a deviation, as a short vowel is expected. As for /u uː/, the values are relatively consistent: the short variant shows relatively higher F1 and F2 values, while the long variant shows relatively lower ones. Vowels /ɪ iː/ behave quite consistently as well.

Table 5 Formant values for the a) SS, b) SL, c) LS, d) LL quantity pattern grouped by vowel quality combinations. S = short vowel, L = long vowel.

a) SS	1st vowel				2nd vowel			
	Hz		ERB		Hz		ERB	
	F1	F2	F1	F2	F1	F2	F1	F2
a - ɪ	824.10	1492.03	14.19	18.75	377.79	2526.79	9.06	23.13
a - u	809.07	1402.47	14.05	18.25	376.40	1048.60	9.04	15.98
ɪ - ɪ	361.14	2565.50	8.80	23.26	344.79	2552.86	8.54	23.21

b) SL	1st vowel				2nd vowel			
	Hz		ERB		Hz		ERB	
	F1	F2	F1	F2	F1	F2	F1	F2
a - i:	872.83	1489.83	14.61	18.74	349.87	2653.52	8.62	23.54
a - u:	853.71	1390.29	14.44	18.18	340.50	948.64	8.47	15.22
ɪ - i:	363.63	2471.88	8.84	22.94	337.25	2650.25	8.42	23.53

c) LS	1st vowel				2nd vowel			
	Hz		ERB		Hz		ERB	
	F1	F2	F1	F2	F1	F2	F1	F2
a: - ɪ	874.09	1430.27	14.62	18.41	386.00	2523.95	9.19	23.12
a: - u	896.38	1415.88	14.80	18.33	374.94	996.25	9.02	15.59
i: - ɪ	343.00	2583.38	8.51	23.32	376.13	2622.50	9.04	23.44

d) LL	1st vowel				2nd vowel			
	Hz		ERB		Hz		ERB	
	F1	F2	F1	F2	F1	F2	F1	F2
a: - i:	871.06	1439.03	14.59	18.46	341.19	2549.29	8.48	23.20
a: - u:	871.38	1415.38	14.59	18.33	347.63	887.25	8.59	14.73
i: - i:	344.33	2589.33	8.53	23.34	340.33	2554.73	8.47	23.22

The short variant tends to have higher F1 and lower F2 values, while the long variant tends to have lower F1 and higher F2 values. However, these differences are relatively small, and particularly for F2, the pattern is less regular, with more deviations from the described tendency.

3.2 Duration measurement results

All values presented below are normalized (see Method). Table 6a shows durations according to the first segmentation variant (i.e., based on formant structure), while Table 6b presents durations according to the second segmentation variant (i.e., based on oscil-

logram). In addition to the mean duration, standard deviation and the ratio between long and short variant of the respective vowel are provided. Vowels in the table are classified as short or long based on the length indicated in the original text, not on perception.

Table 6a Normalized duration of individual vowels, segmentation based on formants (variant 1).

segmentation 1	mean dur. (ms)	SD (ms)	V:/V ratio
a	121.35	34.82	1.36
a:	165.41	42.70	
ɪ	113.88	37.20	1.08
i:	122.56	41.73	
u	111.13	43.77	1.30
u:	144.62	49.62	

Table 6b Normalized duration of individual vowels, segmentation based on oscillogram (variant 2).

segmentation 2	mean dur. (ms)	SD (ms)	V:/V ratio
a	131.03	35.76	1.33
a:	174.39	39.32	
ɪ	130.77	45.83	1.08
i:	141.29	47.27	
u	135.84	49.56	1.27
u:	172.59	53.52	

Although the specific duration values differ for segmentation variants 1 and 2, the ratios between the duration of long and short vowels are very similar for both. A higher ratio can be observed for /a a:/ and /u u:/, and a lower ratio for /ɪ i:/. For this reason, I will continue to present results only according to the first segmentation variant, which follows the established segmentation rules (see Method) and is therefore considered default.

Vowels /ɪ i:/ appeared in both the first and second syllable in the analysed material (see Table 7). The data show that the behaviour of /ɪ i:/ in the first and second syllable differed. In the first syllable, the difference between short and long vowel variants is noticeable, and the long to short vowel duration ratio is comparable to the values in Table 3. However, in the second syllable, short and long vowel variants have the same duration (which corresponds to the duration of long vowel in the first syllable). I was interested in whether these differences would also be reflected in listeners' perception (for more information about the listeners and the perception test see Chabrová & Veroňková, 2022). However, the agreement between listeners' perception and the original text for /ɪ i:/ vowels, considered separately in the first and second syllable, differs only slightly (66% in the first syllable and 59% in the second syllable). The behaviour of /ɪ i:/ according to the position is also noted by Podlipský, Skarnitzl, & Volín (2009), although in their study this concerned the final position of an utterance compared to non-final positions. In the present experiment, the words were never positioned at the edges of sentences.

Table 7 Normalized mean duration of /ɪ i:/ vowels grouped by first and second syllable.

	1st syllable			2nd syllable		
	mean dur. (ms)	SD (ms)	V:/V	mean dur. (ms)	SD (ms)	V:/V
ɪ	87.19	20.81	1.35	121.82	37.30	1.02
i:	117.88	26.35		123.98	45.27	

The results of the duration analysis were further divided according to quantity patterns (SS, SL, LS, LL) and, within each pattern, by vowel quality combinations, as shown in Tables 8a–8d. In order to present all vowel combinations within a single table, I use the labels V1 (first vowel of the word) and V2 (second vowel of the word) instead of specifying the exact vowel quality.

Table 8 Normalized durations of vowels in the a) SS, b) SL, c) LS, d) LL quantity pattern. S = short vowel, L = long vowel.

a) SS	duration V1 (ms)	duration V2 (ms)
a - ɪ	119.21	113.99
a - u	102.71	119.20
ɪ - ɪ	133.49	131.31

b) SL	duration V1 (ms)	duration V2 (ms)
a - i:	133.49	122.50
a - u:	125.81	124.10
ɪ - i:	90.52	95.66

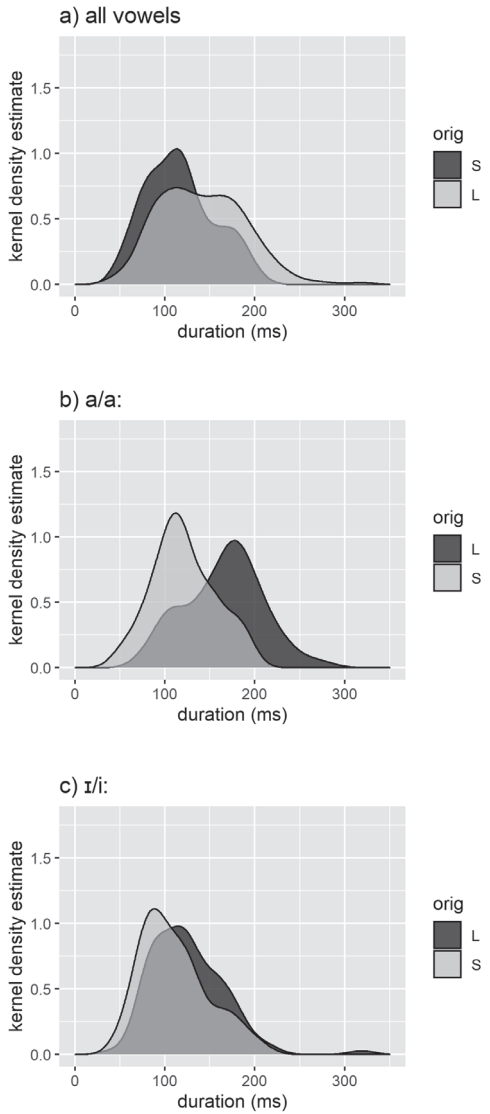
c) LS	duration V1 (ms)	duration V2 (ms)
a: - ɪ	179.88	121.68
a: - u	155.45	108.81
i: - ɪ	140.48	132.08

d) LL	duration V1 (ms)	duration V2 (ms)
a: - i:	169.92	121.43
a: - u:	128.06	161.69
i: - i:	110.31	143.00

The data show that individual vowels behave differently across various quantity patterns (SS, SL, LS, LL) and even within different quality combinations of the same pattern. The vowels /a a:/, which always occur in the first syllable in the analysed words, are longer in all quantity patterns when followed by /ɪ i:/ in the second syllable than when followed by /u u:/ (however, the vowels were not adjacent; there was always a consonant between them). The intended short /a/ is overall longer in the SL pattern than in the SS pattern, and the intended long /a:/ is overall longer in the LS pattern than in the LL pattern.

Vowels /u u:/, which always occur in the second syllable, behave in the opposite manner. Short /u/ is longer in the SS pattern compared to LS, and the long /u:/ is longer in the LL pattern compared to SL. The vowels /ɪ i:/ behave inconsistently, and no clear patterns can be observed.

Figures 2a–2d show a graphical representation of the relationship between vowel duration and whether the vowel was originally supposed to be short or long according to the source text. The graph demonstrates a significant overlap between the areas of short and long vowels, indicating that vowels originally classified as short and long were pronounced with similar durations.



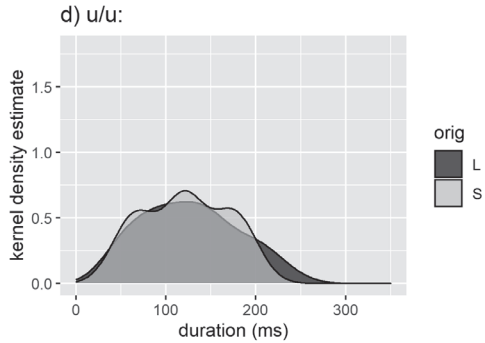
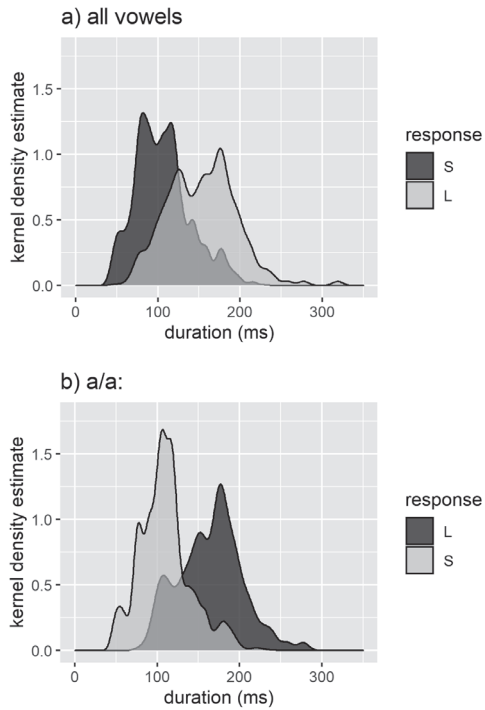


Figure 2 Relationship between normalized vowel duration (ms) of a) all vowels, b) /a aː/, c) /ɪ iː/, d) /u uː/ and the original classification of vowels as short (S) or long (L) according to the source text.

Figures 3a–3d illustrate the relationship between vowel duration and whether the vowel was perceived by the native Czech listeners as short or long. Compared to Figures 2a–2d, a difference is noticeable: the areas for short and long vowels are more distinct, and both variants show clearer peaks. This indicates that listeners were largely guided by duration when perceiving vowel length. However, the S and L areas still overlap to a large extent, which means that vowels of the same duration were often evaluated both as short and long – this could be either due to low inter-listener agreement, or because listeners also relied on other cues, such as vowel quality.



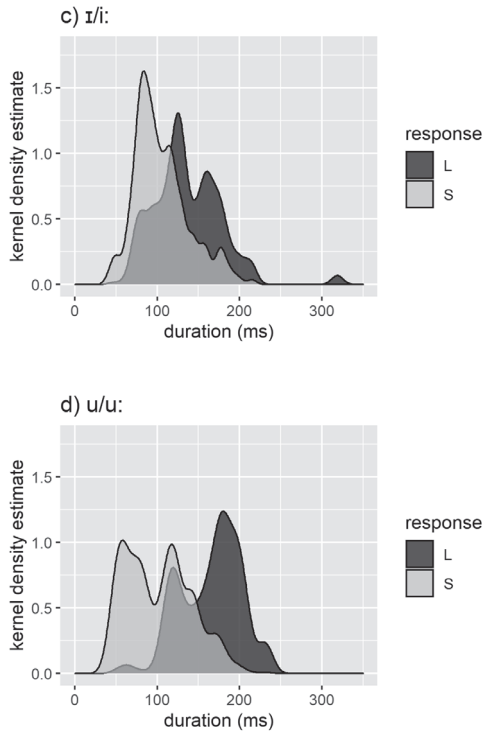


Figure 3 Normalized vowel duration (ms) of a) all vowels, b) /a a:/, c) /ɪ i:/, d) /u u:/ according to whether it was perceived by Czech listeners as short (S) or long (L).

4. Discussion

The previous section presented the results of acoustic analysis of formants and duration. I will now compare my results with reference values. As for the formant measurements, in this experiment I worked with reference data for male speakers published by Skarnitzl & Volín (2012), which I adjusted by increasing the values by 17.5%. At the time the experiment was conducted, this was the best available option. However, I now also have access to reference values for female speakers, which come from the above-mentioned study and have not yet been published, see Table 9. I would like to thank R. Skarnitzl for kindly providing these unpublished reference values (Hz) for female speakers. The conversion of values from Hz to ERB and calculation of the percentage difference between formants of the short and long vowel variants were carried out by the author of this paper; for details on the conversion method, see the Method section.

The comparison of the experimental results (see Table 3) with the reference values above shows that in the analysed material, native speakers of German do not exhibit such large differences in quality between the short and long vowel variants. As for /a a:/, the native speakers of German approximate the reference values well and pronounce the

Table 9 Unpublished F1 and F2 reference values for selected vowels in female Czech speakers, measured in the study by Skarnitzl & Volín (2012), rounded to the nearest ten. The ERB values and F1/F2 difference were calculated by the author of this paper.

	F1 (Hz)	F2 (Hz)	F1 (ERB)	F2 (ERB)	F1 difference (% of ERB)	F2 difference (% of ERB)
a	770	1500	13.70	18.79	1.93	2.40
a:	800	1420	13.97	18.35		
ɪ	490	2250	10.64	22.14	28.19	5.26
i:	330	2600	8.30	23.37		
u	420	1140	9.69	16.62	14.54	22.21
u:	340	760	8.46	13.60		

short and long variants essentially the same. Problems arise with vowels /ɪ iː/ and /u uː/. The results of the present study show that the speakers are going in the right direction when it comes to the realisation of the quality of these vowels – a difference in F1 values can be observed for /ɪ iː/, and in both F1 and F2 values for /u uː/. However, in order for the difference in quality to approach the reference values for native speakers, it would have to be much more pronounced.

An interesting comparison can also be made with the reference values for native German speakers (Sendlmeier & Seebode, n.d.), see Table 10. All six vowel qualities analysed in this study occur in German as well, with the difference that short /u/ is transcribed as /ʊ/ in German, as it reflects a different qualitative character of the sound.

Table 10 F1 and F2 reference values for selected vowels in female native German speakers (Sendlmeier & Seebode, n.d.). The ERB values and F1/F2 difference were calculated by the author of this paper.

	F1 (Hz)	F2 (Hz)	F1 (ERB)	F2 (ERB)	F1 difference (% of ERB)	F2 difference (% of ERB)
a	836	1586	14.29	19.25	3.45	1.91
a:	896	1517	14.8	18.89		
ɪ	433	2095	9.87	21.54	26.21	6.95
i:	302	2533	7.82	23.15		
ʊ	442	1081	10	16.21	17.10	6.09
u:	345	956	8.54	15.28		

When comparing the German and Czech reference values, we see that the vowels /a aː/ generally have higher F1 and F2 formants in German. The vowel /ɪ/ is slightly higher in German, while /iː/ is slightly lower. The vowel /ʊ/ shows only minor differences, and a more pronounced distinction can be observed in the F2 of /uː/. These disparities naturally also affect the F1 and F2 differences between short and long vowels. An important question is therefore: Could the different nature of the target vowels in German have influenced their non-canonical pronunciation in Czech? Based on my data, this can nei-

ther be confirmed nor ruled out. It might be possible for the vowels /a a:/, whose F1 values indeed correspond more closely to the German reference values, but their F2 values, on the other hand, tend to be closer to the Czech reference values. As for the other four vowels, the differences do not seem to be caused by different vowel qualities in the two languages, but rather by the fact that the Czech short vowels /ɪ/ and /u/ are pronounced very similarly to their long counterparts. The only case where the experimental results approach the German reference values is in the F2 of /u:/. Speakers are thus more likely unaware of the differences in vowel quality of these sounds.

As for the absolute vowel duration and the duration ratio between long and short vowels, the results of this study (Table 6a and 6b) can be compared with the reference values presented in the Introduction (Table 1). Whether we take the study by Skarnitzl (2012) or that by Paillereau and Chládková (2019) as a reference, it is clear that the duration differences between short and long vowels are insufficient. Once again, however, it can be observed that the native German speakers are moving in the right direction – the largest duration difference is found for /a a:/, a slightly smaller difference for /u u:/, and only a very small difference for /ɪ i:/, which corresponds primarily to the reference values reported by Skarnitzl (2012). As with vowel quality, it can therefore be concluded that the principle according to which native German speakers in this study distinguish between short and long vowels in Czech words is heading in the right direction, but the vowels need to be differentiated more strongly in order to achieve greater proportional differences.

An interesting comparison can also be made in terms of absolute vowel duration, see Table 11.

Table 11 Duration of native Czech vowels (ms) in Skarnitzl (2012) and Paillereau & Chládková (2019).

	Skarnitzl	Paillereau & Chládková
a	63.1	75
a:	113.0	126
ɪ	53.5	61
i:	68.9	98
u	57.3	74
u:	91.4	119

German speakers in this study produced all vowels with a longer duration than native Czech speakers in the reference studies. This may be because non-native speakers tend to have a slower speech rate, as they require more time to plan their utterances; in the case of reading aloud, they may also read more slowly because they need to concentrate more. In the present study, the average articulation rate of the speakers ranged from 3.29 to 5 syllables per second, with the overall mean articulation rate being 4.20 syllables per second (calculated based on target words, not full sentences). Skarnitzl (2014: 99), who measured speech rate in semi-spontaneous dialogues, found a significant effect of gender in his data, with an average articulation rate of 6.48 syllables per second for female speakers. Different values are reported by Palková (1994: 317–318), who gives a mean

speech rate of 4.98 syllables per second for both genders combined (this value results from averaging five studies, each of which worked with a different type of data). Considering that Skarnitzl worked with articulation rate and Palková with speech rate, the average articulation rates of the speakers in the present study are indeed slightly below average, but the differences are not substantial enough to explain the fact that the vowel durations are sometimes more than twice as long as the reference values.

It is also important to note that the vowels /ɪ iː/ behaved differently in the first and second syllable. Regarding vowel quality, the differences are small, on the order of a few percentage points (the F1 difference is 2.22% lower in the first syllable compared to the second, while the F2 difference is 0.13% higher in the first syllable than in the second; see Tables 5a and 5b). However, for vowel length, the differences are substantial: in the first syllable, the long variant is 35% longer than the short one, whereas in the second syllable, it is only 2% longer (see Table 8). Since /ɪ iː/ occurred more frequently in the second syllable, the overall results show a smaller duration difference between the long and short vowel. The results in Table 8 also suggest that lengthening occurred in the second syllable, as the duration of short /ɪ/ here approximately corresponds to the duration of long /iː/ in both syllables. Since some authors argue that the stressed syllable is relatively longer than the unstressed one in German (see Introduction), I also focused on the absolute duration of vowels in the first syllable compared to the second. However, this hypothesis was not confirmed in my data, see particular Tables 8a–8d for summary results and Table 8 for the vowels /ɪ iː/ in the Results section.

5. Conclusion

This study presented an acoustic analysis of formants and durations of Czech vowels /a aː ɪ iː u uː/ in the speech of native German speakers. Eight native German speakers (all female) were recorded in a studio while reading a text containing sentences with target words. The target words formed groups of two to four words differing only in vowel quantity (e.g., the group of four words platu, platů, plátu, plátů). Formants F1 and F2 were measured, and the differences in formant values between long and short variants of the respective vowels were calculated. The duration of target vowels was also measured, and the extent to which the long vowel variant was longer than the short one was determined. The results were then compared with reference values for Czech and German speakers. The speakers in the present experiment produced correct patterns in terms of how they should distinguish between the long and short vowel variants (differences in quality or duration). However, the differences between long and short vowels were not sufficiently distinct (for example, the vowel qualities /ɪ/ and /iː/ merge together, although they should be two separate vowels, or the duration difference between long /aː/ and short /a/ is not distinct enough). For the vowels /ɪ iː/, which were the only ones found both in the first and second syllables, differences in vowel duration depending on their position in the word were also observed. In the second syllable, lengthening was present, and the duration of both vowels almost completely merged. At the same time, the absolute duration of all vowels was considerably longer than the reference values, to which the speakers' slower articulation rate may have contributed. Differences in formant values when comparing

the results and the reference data for Czech native speakers may or may not be caused or influenced by slightly different formant values for these vowels in the two languages. This study confirmed the assumption that native German speakers tend to struggle with the realisation of vowel length in Czech, and that this phenomenon requires increased attention when acquiring the language.

Acknowledgements

I would like to thank Jitka Veroňková for her valuable consultations during the implementation of the experiment and writing of this article. I also thank Radek Skarnitzl for providing the reference formant values for female native speakers of Czech and Tomáš Bořil for providing initial scripts for the acoustic analysis, which were subsequently adapted for the purposes of this study. This experiment used recordings from the corpus of non-native Czech speech, which was compiled by the Institute of Phonetics at the Faculty of Arts, Charles University, and collected as part of the Czech Science Foundation project GA ČR 18-18300S *Zvukové vlastnosti češtiny v komunikaci nerodilých a rodilých mluvčích*.

REFERENCES

Ashby, M., & Maidment, J. A. (2015). *Úvod do obecné fonetiky*. Karolinum.

Becker, T. (2012). *Einführung in die Phonetik und Phonologie des Deutschen*. WBG Wissenschaftliche Buchgesellschaft.

Boersma, P., & Weenink, D. (2022). *Praat: doing phonetics by computer* (Version 6.2.10). Retrieved 26. 3. 2022 from http://www.praat.org.

Bořil, T. (2015). *Výpočet formantů*. Retrieved 26. 5. 2025 from https://fu.ff.cuni.cz/AKU/script_01.html#1.

Chabrová, A., & Veroňková, J. (2022). Percepce kvantity vokálů v českých dvojslabičných slovech v řeči rodilých mluvčích němčiny. *Nová čeština doma a ve světě*, *2*, 25–37.

Glasberg, B. R., & Moore, B. C. J. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing research*, *47*(1), 103–138.

Jessen, M. (1993). Stress conditions on vowel quality and quantity in German. *Working Papers of the Cornell Phonetics Laboratory*, *8*, 1–27.

Kleiner, S., Knöbl, R. (2015). *Duden, das Aussprachewörterbuch* (7th ed.). Dudenverlag.

Křen, M., Cvrček, V., Henyš, J., Hnátková, M., Jelínek, T., Kocek, J., Kováříková, D., Křivan, J., Milička, J., Petkevič, V., Procházka, P., Skoumalová, H., Šindlerová, J., & Škrabal, M. (2020). *SYN2020: reprezentativní korpus psané češtiny*. Ústav Českého národního korpusu FF UK. Retrieved from http://www.korpus.cz.

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

Paillereau, N., & Chládková, K. (2019). Spectral and temporal characteristics of Czech vowels in spontaneous speech. *Acta Universitatis Carolinae. Philologica*, *2019*(2), 77–95.

Palková, Z. (1994). *Fonetika a fonologie češtiny: s obecným úvodem do problematiky oboru*. Karolinum.

Palková, Z., Bořil, T., & Veroňková, J. (2020). Difficulties in adjacent vowel length of L1 Russian speakers in Czech. In Botinis, A. (ed.), *Proceedings of 11th International Conference of Experimental Linguistics* (pp. 149–152). ExLing.

Podlipský, V. J., Skarnitzl, R., & Volín, J. (2009). High front vowels in Czech: a contrast in quantity or quality? In *Proceedings Interspeech 2009* (pp. 132–135).

Pompino-Marschall, B. (2003). *Einführung in die Phonetik* (2nd ed.). Walter de Gruyter.

Rausch, R., & Rausch, I. (1998). *Deutsche Phonetik für Ausländer* (5th ed.). Langenscheidt.
Sendelmeier, W. F., & Seebode, J. (n.d.). *Formantkarten des deutschen Vokalsystems*. Retrieved 25. 3. 2022 from https://www.tu.berlin/kw/forschung/projekte/formantkarten.
Skarnitzl, R. (2012). Dvojí i v české výslovnosti. *Naše řeč*, *95*(3), 141–153.
Skarnitzl, R. (2014). *Fonetická identifikace mluvčího*. Filozofická fakulta Univerzity Karlovy.
Skarnitzl, R. (2018). Fonetická realizace slovního přízvuku u delších slov v češtině. *Slovo a slovesnost*, 79(3), 199–216.
Skarnitzl, R., & Volín, J. (2012). Referenční hodnoty vokalických formantů pro mladé dospělé mluvčí standardní češtiny. *Akustické listy*, *18*(1), 7–11.
Vachek, J. (1968). *Dynamika fonologického systému současné spisovné češtiny*. Academia, nakladatelství Československé akademie věd.
Veroňková, J., & Bořil, T. (2020a). Czech vowel quantity in Polish speakers as perceived by Moravian-Silesian listeners. In *Speech Research conference* (pp. 89–91). Hungarian Research Institute for Linguistics.
Veroňková, J., & Bořil, T. (2020b). Phonological length of L2 Czech speakers' vowels in ambiguous contexts as perceived by L1 listeners. In Karpov, A., & Potapova, R. (eds.), *Speech and Computer. 22nd International Conference, SPECOM 2020*. *Lecture Notes in Computer Science*, vol. 12335 (pp. 624–635). Springer.
Veroňková, J., Bořil, T., Palková, Z., & Poukarová, P. (2020). Délka českých samohlásek u polských mluvčích v taktech s různou strukturou kvantity. In Bogoczová, I. (ed.), *Area Slavica 3 (Jazyk na hranici – hranice v jazyku)* (pp. 51–61). Ostravská Univerzita, Filozofická fakulta.
Wiese, R. (2011). *Phonetik und Phonologie*. Wilhelm Fink.

RESUMÉ

Předkládaná studie je součástí širšího experimentu, který se zaměřuje na percepční a akustickou analýzu vokálů /a a: ɪ i: u u:/ v českých dvojslabičných slovech u rodilých mluvčích němčiny, učících se češtinu jako cizí jazyk. Studie prezentuje výsledky akustické analýzy a poskytuje stručný přehled o chování vokálů v češtině a němčině. Materiál tvoří dvojice, trojice a čtveřice slov, které se navzájem liší pouze kvantitou vokálu (např. čtveřice platu, platů, plátu, plátů). Nahrávky pochází od 8 žen různých jazykových úrovní (A2–B1 podle SERR). V cílových slovech byly analyzovány vokalické formanty F1 a F2 a vypočteny rozdíly ve formantových hodnotách mezi krátkými a dlouhými vokály. Dále bylo změřeno trvání vokálů a byly vypočteno, o kolik jsou dlouhé vokály delší než krátké. Výsledky studie jsou následně porovnány s referenčními hodnotami pro ženské rodilé mluvčí češtiny a ukazují, že mluvčí v předkládané studii nerozlišují krátké a dlouhé vokály dostatečně silně, a to jak kvalitativně, tak kvantitativně, obecně ale volí správné strategie. V oblasti trvání jsou největší rozdíly mezi trváním krátkého a dlouhého vokálu u dvojice /a: a/, následuje /u: u/ a nejmenší rozdíl je u /ɪ i:/. Také z hlediska formantů dochází správně k odlišování rozdílné kvality u dlouhého a krátkého /ɪ i:/, v menší míře také /u u:/. Například u vokálů /ɪ i:/ by ale podle referenční studie měly být hodnoty F1 u dlouhého vokálu o 28 % nižší než u krátkého, zatímco u mluvčích v této studii byl rozdíl pouze 5 %. Hodnoty vokalických formantů u mluvčích v této studií mohou, ale nemusí být ovlivněny mírně odlišnými hodnotami formantů v němčině ve srovnání s češtinou. Výsledky také ukazují, že absolutní trvání vokálů je delší, než uvádí referenční hodnoty pro české mluvčí, což by mohlo být způsobeno nižším artikulačním tempem. Studie potvrdila předpoklad, že vokalická kvantita je jev, kterému by rodilí mluvčí němčiny při osvojování češtiny měli věnovat zvýšenou pozornost.

Anna Chabrová
Institute of Phonetics
Charles University, Faculty of Arts
Prague, Czech Republic
anna.chabrova@ff.cuni.cz

2025 ACTA UNIVERSITATIS CAROLINAE PAG. 177–194
PHILOLOGICA 3

TIMING THE DIFFERENCE: A STUDY OF GEMINATION IN DOGRI CONSONANTS

PRANAV BADYAL

ABSTRACT
This study examines the phonetic and phonological properties of singleton–geminate contrasts in Dogri within the context where the preceding vowel is short, i.e. in CV̲CV: word structures. Linear Mixed-Effects (LMER) statistical results reveal that geminates are temporally nearly twice as long as their singleton counterparts (~ 70 ms), followed by consistent shortening of the preceding vowel (~ 16 ms). The remaining segments, i.e., the word-initial consonant and the word-final vowel, remain indiscriminable across both word types, suggesting no anticipatory lengthening in geminate contexts for the former, and minor shortening (but statistically non-significant) among geminates for the latter.
Additional findings regarding voicing effects, phrase-level context, and overall word duration for both word types confirm cross-linguistic voicing trends: voiceless geminates are longer with added evidence of pre-consonantal vowel shortening relative to their voiced counterparts. Moreover, geminates show significantly longer durations in Fixed Phrase contexts than in Carrier Phrase. Lastly, the longer total word duration for geminates (~ 53 ms) suggests that gemination in Dogri extends beyond a local phonetic phenomenon, bearing a distinctive prosodic and phonological weight. These results posit the view that temporal elongation reflects their status as phonologically contrastive units, rather than being a by-product of durational redistribution of the segments.

Keywords: Dogri; geminates; singleton; Punjabi; gemination; duration; temporal; Jammu; Kashmir; Indo Aryan; length contrast; tenseness

1. Introduction

The western Pahari sub-branch of the modern Indo-Aryan language, Dogri (Glottocode: dogr1253, ISO 639-3: doi) is primarily spoken in the Jammu and Kashmir (J&K) region of India. Its origin can be traced through earlier historical phases, namely the Old Indo-Aryan (OIA) period (approximately 1500 B.C. to 600 B.C.) and the Middle Indo-Aryan (MIA) period (approximately 600 B.C. to 1000 A.D.) (Gupta & Chowdhary, 2024). It is spoken by approximately 0.21% of the Indian population (Language Census of India, 2011). As there are around 60,000 people who belong to diverse religions and tribes, who still practice transhumance around the J&K region (Warikoo, 2000), Dogri is

https://doi.org/10.14712/24646830.2025.27

sparingly spoken in the adjoining areas of Himachal Pradesh state and the northern areas of Punjab state. From a script-historical perspective, Dogri has undergone three phases of script development: Takri, New Dogra Script, and Devanagari. It is currently written in the Devanagari script, the same script used for Hindi, which itself evolved from the Sharada script around 1200 A.D. (see Gupta & Chowdhary, 2024, p. 8; Grierson, 1904, p. 67 for Takri script).

1.1 Purpose of the study

This paper aims to investigate the durational properties of singleton and geminate consonants in Dogri. To contextualise the findings, this study will draw cross-linguistic comparisons with related Indic languages and typologically diverse languages that exhibit a short–long consonant contrast.

However, given the limited body of linguistic research on Dogri in general, certain considerations are crucial in shedding more light on gemination patterns. For example, in a pilot study, Badyal (2023) focused solely on five voiceless stops [p, t̪, ʈ, tʃ, k], thus offering a narrow phonetic scope. Concerning Ghai (1991), it presents some limitations: some test words are either nonsense (made-up) or borrowed from Punjabi, possibly due to the author's native background Also, the analysis was performed manually, with spectral data presented as mingograms derived from tape-recorded acoustic signals.

By addressing the questions on Dogri geminates below, the study aims to confirm, refine, or potentially revise earlier findings, accounting for variation in the form of language change in Dogri, if any, particularly after 1991, by addressing the following questions:

a. C2 duration difference between singleton and geminate words.
b. Effect of place and voicing on C2 duration across both word types.
c. Effect of C2 on V1 and V2 duration in singleton vs. geminate contexts.
d. Are voiceless geminates longer than voiced ones?
e. Evidence of pre-consonantal vowel (V1) duration before voiced compared to voiceless for both word types.
f. Do geminate durations vary between fixed and carrier sentence contexts?
g. Is there C1 lengthening in geminate constructions?
h. Is the total word duration preserved between singleton and geminate words?

1.2 Geminates in Dogri

A geminate, usually defined as one single unit or a sequence of two identical consonants, is etymologically derived from a Latin word, 'geminus', meaning '*doubling*'. Most languages in the world have short/single consonants in their inventory, while some maintain a two-way length, i.e., a phonological quantity contrast in the form of a singleton-geminate that shows meaning contrast. This contrast offers a distinction between weak or lenis–strong or fortis–tense or lax consonant system in such languages.

As Blevins (2004, 2005) suggests, geminates may emerge through multiple evolutionary pathways including the assimilation of consonant clusters (CC) (see Ohala, 2007 for CC assimilation in Hindi), consonant + vowel/glide sequences, expressive or emphatic

lengthening, lengthening at word or morpheme boundaries, and occasionally through language contact (e.g., the emergence of voiced geminates in Japanese via loanwords), a comprehensive discussion of these origins falls beyond the scope of this paper. Thus, by limiting to Dogri and languages of Indic origin, in addition to native lexical geminates, the rise of geminates is closely linked with the historical assimilation of two clustered consonants (CC) among disyllable *tatsam* (of Sanskrit origin) words. It is triggered when one consonant occupies the coda of the preceding syllable and the other the onset of the following syllable. In case of word-final geminates, a geminate is formed when both the CC members are in the coda of simple monosyllable words (cf. Table 1 for the origin of Dogri geminates from CC).

Table 1 Geminates arising from CC assimilation in Dogri.

POA	Sanskrit	Hindi	Dogri	Meaning
palatal + retroflex	əʃʈ	ɑːʈʰ	əʈʰː	eight
labial + dental	səpt̪	sɑːt̪	sət̪ː	seven
alveolar + dental	həst̪	hɑːt̪ʰ	ət̪ʰː	hand
alveolar + labial	sərp(ə)	sɑ̃ːmp	səpː	snake
velar + palatal	sə.məkʃ	sə.məkʃ	sə.məkʰː	in front
velar + palatal	d̪ək.ʃɪn	d̪ək.ʃɪn	d̪əkkʰən	south
dental + alveolar	suːt̪r	suːt̪r	sʊt̪ːər	thread
dental + alveolar	muːt̪r	muːt̪r	mʊt̪ːər	urine
dental + alveolar	pɔːt̪r	pʊt̪r	pʊt̪ːər	son
dental + alveolar	kʃeːt̪r	kʃeːt̪r	kʰeːt̪ːər	an area

Dogri is a phonetic language (words are spelt as they are written, cf. Table 2 for [d̪] minimal pairs with short and long pregeminate vowel examples) that allows 19 consonants that can appear as geminates in Dogri in the word-medial and word-final position. Consonants including stops [p, pʰ, t̪, t̪ʰ, ʈ, ʈʰ, k, kʰ, d̪, ɖ, b, g], palatals [tʃ, tʃʰ, dʒ], sonorants [m, n, l], and a fricative [s] can appear as geminates in Dogri. All these geminate consonants also have a singleton form. Consonants [r, ɳ, ɽ, ʃ] only appear as singletons (Kaur & Dwivedi, 2018). An absolute word-final geminate in Dogri is a closed syllable that occurs after a short vowel only, and the geminate is followed by a non-phonemic vocalic release. Word-initial geminates are not allowed. Geminates do not occur flanked by another consonant on either side in a word in Dogri.

With respect to the preceding vowel in a CVCV template, geminates can occur in two types of environments across languages:

where a geminate appears immediately after a short-stressed vowel only, as in Swedish, Italian, Icelandic, Luganda, Hindi, Punjabi (Dulai & Koul, 1980) and,

where a geminate can appear both after a short as well as a long vowel, i.e., after central–peripheral, like in Dogri (see Table 2 below for examples), Lebanese Arabic (Khattab & Tamimi, 2014, p. 238) and Ta'zi dialect of Yemeni Arabic (Aldubai, 2015, p. 341).

Table 2 Orthographical representation of the Dogri singleton–geminate pair in IPA and Devanagari script.

Word Pair	IPA (Singleton)	IPA (Geminate)	Devanagari (Singleton)	Devanagari (Geminate)	Meaning (Singleton)	Meaning (Geminate)
1	səɖɑː	səɖɖɑː	सदा	सद्दा	always (verb, simple present)	call (verb, continuous)
2	dʒɑːɖiː	dʒɑːɖɖiː	जादी	जाद्दी	more	freedom

Languages that permit geminates in the context described in 2(b) are relatively rare. Many scholars have proposed that a short stressed centralised vowel preceding the word-medial consonant constitutes the most favourable phonetic environment for the realisation of geminates in Indic languages. Punjabi, the closest language to Dogri in the Indic group, adheres to this rule (see Bhatia, 1993; Gill & Gleason, 2013, p. 22; Hussain, 2015; Maddieson, 1985, p. 212). This preference may be attributed to the fact that a short vowel in a first stressed syllable creates an articulatory setting that facilitates sufficient constriction or frication, enabling the following consonant to be perceived and produced as durationally longer than its singleton counterpart. As a result, in Dogri, minimal word pairs containing a short vowel before the geminate consonant (CVCV) are more commonly attested than those featuring a long vowel in the same position (CV:CV) when contrasting singleton and geminate (S–G) forms. By word template, Dogri geminates are presented in Table 3.

Table 3 Dogri geminates occurring in different word templates among short and long pregeminate vowels.

Word Template	Word (IPA)	Preceding Vowel	Geminates in Word	Meaning
CVC:	sʊʈʈ	Short	ʈʈ	throw (imperative)
CVC:V:	ʧəkkɑː	Short	kk	wheel
CV:C:V:	ʧɑːkkiː	Long	kk	soap
CVC:V:C	kʰəɖɖəɽ	Short	ɖɖ	a kind of cloth
V:C:VC	ɑː.kkʰən	Long	kkʰ	ask (imperative)
CVCVC:V:	ʧə.kənn.ɑː	Short	nn	watchful
CVCV:C:V:	ʧə.lɑːkk.iː	Long	kk	cleverness

1.3 Temporal correlates of geminates

The study will focus on the temporal distribution of four segments, i.e., CVCV, in a S–G minimal pair across previously studied languages. It will also be examined how languages differ in the temporal distribution among segments. However, efforts will be made consistently to group findings that are consistent with Dogri and related Indic languages.

Within the CVCV minimal pair word structure, the absolute duration, comprising the closure and release phase of the intervocalic consonant (C2) stops, is the primary

acoustic correlate distinguishing singletons from geminates. This durational contrast is widely considered a universal phonetic phenomenon across languages. As Ladefoged & Maddieson (1996, p. 92) find, the absolute duration of C2 between singleton and geminates varies cross-linguistically. The evidence from previous research shows that across a range of typologically diverse languages, an S–G duration ratio of approximately 1:2 is commonly observed. Languages that adhere to this ratio include Dogri (Badyal, 2023), Cypriot Greek (Arvaniti, 1999), Yemeni Arabic (Aldubai, 2015) and Italian, specifically for stop consonants [p, t, k] (Esposito & Di Benedetto, 1999). The duration ratio higher than 1:2 however is attested in languages like Hindi, 1:2.5 (Shrotriya et al., 1995; Ohala, 2007, p. 354), Polish, 1:2.48 (Rojczyk, 2019), Pattani Malay (Abramson, 1986), while Berber languages demonstrate an overall consonant duration ratio of approximately 1:3 (Khattab & Tamimi, 2014, p. 232) (also see Table (ii) in Hamzah et al., 2016 for a detailed review on S–G duration ratio). However, in comparative studies, the consistency of these duration ratios remains a topic of debate due to various influencing factors, such as variability in speech rate (see Mitterer, 2018), positional effects (e.g., whether the target word is utterance-initial or phrase-medial), communicative context, and the semantic relation of the word within the carrier phrase, etc.

Regarding voicing of C2 (voiced vs. voiceless) among geminates, C2 voicing has a strong effect on its own duration, with voiceless consonants generally being longer than voiced ones. Importantly, the voicing of C2 also influences the duration of the preceding vowel (V1); voiceless geminates tend to be longer in duration than voiced ones. Such observations are made for languages like Hindi (Samudravijaya, 2003; Shrotriya et al., 1995, p. 133). In terms of articulatory phonetics, voiceless consonants require more precise control of airflow and greater tension in the articulators, which leads to a longer closure period. In contrast, voiced geminates require activation of vocal fold vibration and less airflow and tension, which lead to shorter durations.

Although geminates universally show compensatory lengthening of the C2, they also trigger temporal redistribution for the rest of the segments in a CVCV word template. The second important correlate in distinguishing geminates from singleton lies in the duration of the flanking vowels, namely, the preceding vowel (V1) and the following vowel (V2) relative to the intervocalic consonant (C2). With respect to V1 timing, Ridouane (2010) reports that in at least nine languages, the length of V1 plays a significant role in geminate contrasts. In the context of Indian languages, V1 shortening in geminate environments is a widely observed phenomenon, for example, in Hindi (Ohala, 2007, p. 355; Shrotriya et al., 1995, p. 134), Bengali (Lahiri & Hankamer, 1999), and Dogri (Badyal, 2023; Ghai, 1991). Previous research on V1 duration among Dogri geminates records a reduction of an average of 17 ms (Badyal, 2023). Interestingly, because geminates primarily emerged as a result of CC assimilation in Indic languages, similar reductions in the vowel preceding the cluster have also been reported, for example, in the case of Hindi (Shrotriya et al., 1996). With regard to V1 duration among voiced–voiceless geminates, it is also found that V1 duration tends to be longer among voiced stops than in comparison to voiceless stops (see Samudravijaya, 2003 for Hindi). This phenomenon is evident even in simple monosyllable words in non-geminating languages like English; consider word pairs, *bead–beat, seed–seat*, a range of 0.6 ~ 0.8 ms is attested (Cho, 2016; also see Kluender et al., 1988 for a review on production-oriented vowel length effect).

Concerning V2 duration, previous research shows that language-specific rules govern the role of the following vowel in differentiating singleton and geminates. While for certain Indic languages, notably Dogri (Badyal, 2023), and specifically Bengali, Ghosh (2015), in a three-way ANOVA examines the effects of gemination, place of articulation, and voicing on V2 duration and finds no statistically significant difference between singletons and geminates. Japanese, however, maintains a shorter duration for V2 when they follow geminates (Han, 1994, also cited in Hirata, 2007, p. 10). Similar V2 shortening effects have also been observed in Pakistani Punjabi (Hussain, 2015). However, findings from the Ta'zi dialect of Yemeni Arabic demonstrate a more consistent pattern, i.e., both V1 and V2 durations are reduced in geminate environments (Aldubai, 2015). These cross-linguistic differences suggest that V2 durational patterns in geminate contexts are language-specific, reflecting variation in phonetic realisation and prosodic structuring across languages.

Robust C1 lengthening effects have been reported for geminates and word-medial CC in languages such as Japanese (Han, 1994), Yemeni Arabic (Aldubai, 2015), and Italian (Turco & Braun, 2016). Aldubai's study on Yemeni Arabic, in particular, demonstrates C1 nasal [m] in geminate environment to be more than twice as long as in singleton contexts, ranging between 60 and 150 ms. The presence of C1 lengthening in geminates (and/or consonant clusters) contexts, whether consistent or variable across languages, is considered a natural and physiologically grounded phenomenon. It may be performed to facilitate the production of a longer C2 by providing articulatory strengthening or preparatory support during the utterance of a geminate word. However, in contrast, minor (non-significant and less consistent) lengthening effects have been observed in Pakistani Punjabi (Hussain, 2015) and in Hindi (Ohala, 2007, p. 357).

As discussed in the above section on the temporal distribution of segments in S–G word pairs, languages demonstrate distinct timing patterns, reflecting language-specific strategies that account for cross-linguistic variation between the two word types. Despite this variability, it is still possible to classify languages based on the durational characteristics of the pregeminate vowel because it emerges as the second most widely studied and is considered a reliable correlate in terms of reflecting significant statistical results across speakers and contexts, particularly in support of Indo-Aryan languages spoken in India. In this respect, languages can be grouped on the basis of three types:

a. Languages like Hindi (Ohala, 2007; Samudravijaya, 2003; Shrotriya et al., 1995), Bengali (Ghosh, 2015; Lahiri & Hankamer, 1988), Dogri (Badyal, 2023; Ghai, 1991, p. 38), Italian (see Esposito & Benedetto, 1999, p. 2059), Makasar (Tabain & Jukes, 2016, p. 103) and Tashlhiyt Berber (Ridouane, 2007) are reported to show the preceding vowel to be shorter among geminates in comparison to singleton in a CVCV minimal word pair.
b. Languages that lengthen the preceding vowel before a geminate, most notably in Japanese (Han, 1994; Hirata, 2007; Idemaru & Guion, 2008).
c. Languages that are reported to show no reductions in the duration of the preceding vowel between singletons and geminates are Polish (Rojczyk & Porzuczek, 2019), Maltese (Mitterer, 2018), Pakistani Punjabi (Hussain, 2015), Hungarian (Ham, 2001), Estonian (Engstrand & Krull, 1994) and Turkish (Lahiri & Hankamer, 1988).

2. Methods

2.1 Participants

Ten participants (6 males and 4 females) with no reported history of speech or language disorders and between the ages of 25 and 69 were recruited in Jammu, J&K, India. All participants were at least bachelor's level or higher educated and had been residing in the Jammu region for a minimum of ten years, and none had resided outside of India. In terms of competence in other languages, all participants reported fluent articulation in Hindi and English (English language onset age: no later than 8 years). Each participant reported having learned Dogri prior to English and used Dogri for at least 70% of their daily communication.[1]

2.2 Materials and experimental design

The material consisted of 13 pairs of disyllabic test words following a CVCV word template, each forming a true minimal pair contrasting singleton and geminate consonants. The word-medial consonants were selected based on four places of articulation: (i) bilabial, (ii) dental, (iii) retroflex, and (iv) velar, further categorised according to the manner of articulation (stop, nasal, liquid) and also by phonological voicing (voiced–voiceless). Six voiceless C2 were included, whereas voiced consonants were seven. The word-final vowel in all test word pairs was always long, to control for post-consonantal vowel effects.

Each pair in both environments was embedded in two types of reading stimuli: (i) a fixed carrier phrase, and (ii) meaningful, semantically coherent carrier sentences, and the target word appeared in phrase/sentence-medial position. For the Fixed Phrase condition, the structure [ʊs ____ sʊneːɑː], *he/she* _____ *listen* (simple past) was used, with the blank space representing either a singleton or geminate word. In the meaningful sentence condition, the test words were contextually integrated into natural sentences to preserve the spontaneity and naturalness of speech. This setup was designed to examine whether semantic relatedness influences the articulation of geminates. Filter tokens, i.e. distractor items not analysed in the present study, were also included to avoid the emergence of identifiable patterns. Table 4 presents a full inventory of phonemes included in the stimuli (see Appendix for the reading material).

Table 4 The phonetic distribution of segments in the CVCV word form included in the stimuli.

	Phoneme
No. of word pairs	13
C1	t̪ k b g f tʃ s m r
V1	ə ʊ
C2	p t̪ ʈ ʈʰ k d̪ g l m n
V2	eː ɑː iː

[1] Before participation, all individuals completed an informed consent form in accordance with ethical research procedures.

2.3 Procedure

The recording sessions took place in a sound-attenuated recording studio. All materials were presented in Devanagari script. Before the experiment, the participants were briefly familiarised with the sentence lists, followed by on-screen instructions. Participants were instructed to read at a normal, relaxed speaking rate to minimise hyperarticulation. Participants were allowed to repeat a sentence if they stumbled or mispronounced a word, ensuring fluency and consistency of data.

Each participant was individually recorded reading each test word three times under each stimulus condition, resulting in randomised lists of items. On average, the total duration of the experiment for each participant lasted 10–12 minutes.

All speech was recorded in stereo mode, at 44.1 kHz, and encoded in 16-bit, uncompressed WAV format using a Zoom H1N recorder with cardioid dynamic features. The audio signal was recorded on the hard disk of a desktop computer. Stereo recordings were converted to mono by averaging the two channels, using the default mono conversion feature of the Praat speech software package, version 6.3.16 (Boersma, 2023).

2.4 Measurement and acoustic analysis

Acoustically recorded speech data were annotated for each phoneme of interest using Praat software (see Figures 1 and 2 for the segmentation criteria followed).

The word of interest had the following phases:

a. duration of the word-initial consonant (C1),
b. duration of the preceding vowel (V1),
c. absolute (closure and release) duration of the word-medial consonant (C2),
d. duration of the word-final vowel (V2), and
e. total duration of the test word (C1+V1+C2+V2 durations)

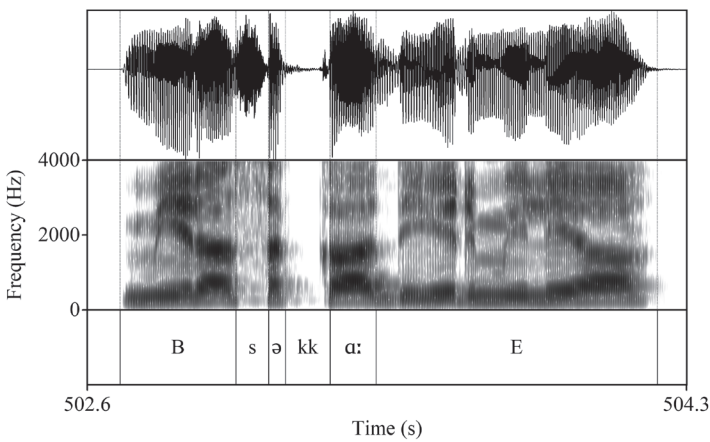


Figure 1 Segmentation of voiceless geminate test word [səkkɑː] in meaningful carrier phrase.

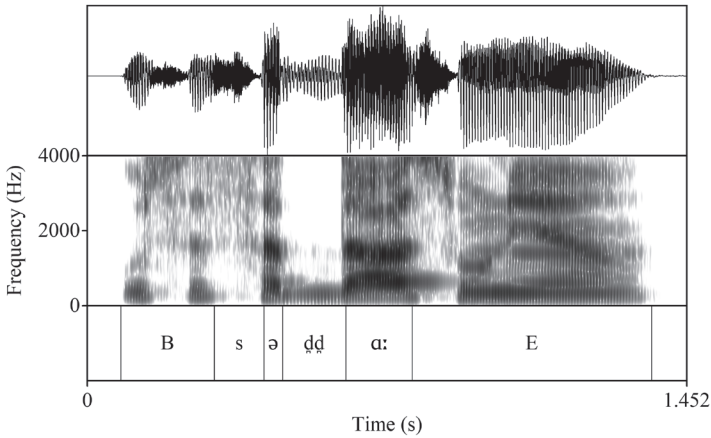


Figure 2 Segmentation of voiced geminate test word [səd̪d̪ɑː] in fixed phrase.

Out of a total of 1560 tokens generated during the experiment (26 words from 13 minimal pairs × 6 repetitions × 10 participants), 770 tokens occurred within the Fixed Phrase environment and 790 within the Meaningful Carrier Phrase.

3. Analyses and results

A linear mixed-effects analysis (LMER) was performed using R and the *lme4* package to examine the relationship between C1–V1–C2–V2 and Consonant Type (singleton/geminate) (Bates et al., 2015). The model included the Consonant Type and Place of Articulation as fixed effects, along with their interaction. As random effects, subject-level intercepts and by-subject random slopes for consonant type were incorporated. Visual inspection of the residual plots revealed no substantial violations of homoscedasticity or normality. Using the *lmerTest* package (Kuznetsova et al., 2017), *p* values were derived via the Satterthwaite approximation. Post hoc analyses (Tukey test) were obtained from the *emmeans* package to draw contrasts between variables (Lenth, 2025).

For C2, the model revealed that there was a significant effect of Consonant Type (S/G) on C2 duration, $F(1, 8.99) = 180.84$, $p < .001$, indicating that geminates were significantly longer than singleton with an estimated mean difference of 72.4 ms. There was also a significant main effect of Place, $F(3, 1542.84) = 27.6$, $p < .001$, and a significant interaction between Consonant Type and Place, $F(3, 1548) = 6.2$, $p < .001$, indicating singleton–geminate contrast varied at four places of articulation.

Concerning V1, a strong effect was found when the word was a singleton or a geminate, $F(1, 9.01) = 23.67$, $p < .001$, maintaining an average duration difference of 16 ms. There was also reported a significant effect of Place, meaning the duration differed between the two word types, $F(3, 1542.20) = 3.89$, $p = .008$. As far as interaction between Consonant Type and Place is concerned, it was found significant, $F(3, 1539.15) = 2.99$, $p = .02$, confirming pre-consonantal vowel shortening among geminates.

V2 had a significant effect of Consonant Type, $F(1, 8.9) = 16.62$, $p < .002$ and Place, $F(3, 1544.89) = 12.24$, $p < .001$, but there was no interaction effect, $F(3, 1541.11) = 1.02$, $p = .3$. The results from pairwise contrasts revealed that despite minute overall shortening of the vowel among geminates, V2 statistically remained similar across all places for singleton and geminates, except for Dentals (7.4 ms), with vowel preceding singleton being longer.

Lastly, for C1, the main effect of Consonant Type was not significant, $F(1, 943.16) = 0.40$, $p = .5$, while Place showed a strong effect, $F(3, 940.11) = 231.28$, $p < .001$. Importantly, there was a significant Word Type × Place interaction, $F(3, 937.10) = 3.70$, $p = .01$, suggesting that the C1 duration contrast between singletons and geminates varied by Place. However, post-hoc comparisons revealed no significant durational differences between singleton and geminate C1 durations at any articulation place. The independent mean duration for four segments grouped by Place of C2 articulation in singleton-geminate word pairs is presented in Table 5.

Table 5 Mean durations for C1–V1–C2–V2 in singleton–geminate CVCV words across four places of articulation. The values in the parentheses represent the standard deviation.

Pair	Place	Voicing	C1 (S)	V1 (S)	C2 (S)	V2 (S)	C1 (G)	V1 (G)	C2 (G)	V2 (G)	Word (S)	Word (G)
p pp	Bilabial	VL	129 (36)	66 (16)	89 (15)	128 (32)	130 (35)	54 (15)	159 (29)	123 (30)	412	465
t̪ t̪t̪	Dental	VL	128 (32)	60 (15)	91 (18)	130 (30)	131 (44)	46 (13)	170 (31)	129 (22)	409	475
ʈ ʈʈ	Retro	VL	123 (27)	69 (17)	81 (19)	127 (29)	132 (34)	54 (14)	149 (18)	118 (28)	400	452
t̪ʰ t̪t̪ʰ	Dental	VL	118 (34)	75 (17)	112 (23)	129 (30)	120 (41)	59 (15)	188 (28)	123 (32)	434	488
k kk	Velar	VL	122 (25)	63 (13)	84 (18)	131 (34)	125 (27)	48 (13)	163 (26)	126 (24)	399	461
d̪ d̪d̪	Dental	V	115 (28)	77 (18)	58 (12)	142 (27)	120 (29)	61 (14)	133 (23)	133 (27)	391	447
g gg	Velar	V	100 (22)	81 (19)	57 (11)	132 (27)	116 (36)	66 (15)	121 (23)	130 (23)	369	433
l ll	Dental	V	126 (39)	73 (20)	51 (10)	136 (30)	133 (44)	55 (17)	125 (24)	125 (26)	385	438
m mm	Bilabial	V	55 (30)	84 (22)	84 (13)	121 (26)	50 (28)	66 (14)	136 (27)	114 (32)	344	365
n nn	Dental	V	130 (35)	78 (19)	48 (20)	127 (37)	126 (35)	54 (15)	133 (30)	121 (26)	382	434

Thus, C2 and V1 provide robust temporal correlates in differentiating Dogri singleton and geminates. As anticipated, at four places of articulation, all 13 phonemes show a 2:1 G–S ratio, singleton maintaining an average of 72 ms shorter duration. The C2 timing

contrast between the two word types by Place type is 64 ms for Bilabials, 78 ms for Dentals, 66 ms for Retroflex, and 72 ms for Velar consonants. Hindi, in contrast, an Indic language, is attested to maintain a larger value of approximately 2.5 times (Shrotriya et al., 1995).

In parallel, V1 is consistently shorter in duration for geminates at all four places of C2 articulation, averaging about 16 ms. This pattern is consistent with an earlier study conducted on Dogri voiceless geminates that showed an average duration difference of 17 ms (Badyal, 2023). The small variation (~ 1 ms) is likely due to natural speech variability and the inclusion of voiced consonants. Similar patterns of pre-consonantal vowel reduction have also been observed in Bengali (Ghosh, 2015), supporting the cross-linguistic preceding vowel shortening generalisation. At four places, namely, Bilabial, Dental, Retroflex, and Velar geminates show an average shortening of 13.5, 18, 15, and 14 ms, respectively, in Dogri. The correlation between V1 and C2 durations by Place of Articulation is presented in Figure 3.

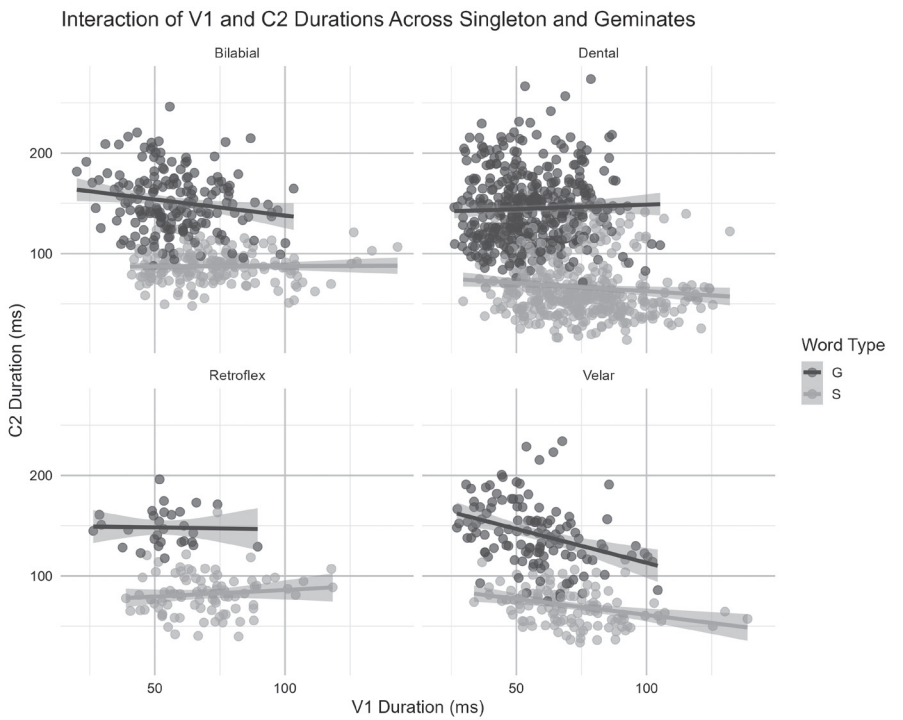


Figure 3 Scatterplot of V1 and C2 durations by word type and consonant place.

Regarding C1, in total, six consonants comprising of four voiced consonants, stops [b, g] and sonorants [m, r], and two voiceless fricatives [f], [s] were included for measuring the temporal difference. These consonants exhibit clearly identifiable onset and offset boundaries due to their manner (voicing for stops) and frication noise (for fricatives). Voiceless stops, however, were not included due to the difficulty in accurately determining

their onset timing. Average durations for each phoneme show that overall, C1 durations show minute but non-significant lengthening effects among geminates for [b, g, f, s], as C1, [b] maintains the longest closure duration due to its feature as voiced bilabial plosive (closure and release burst), while [s] (continuous voiceless frication) the least (cf. Figure 4). However, it is interesting that besides minute lengthening of the consonant among geminates, the C1 in word pairs [rəmiː – rəmmiː] and [məniː – mənniː] is not lengthened, but gets shortened. One potential rationale is that phonetically, sonorants, [m] and [r], generally exhibit shorter durations, with less stable articulatory gestures. Phonologically, in particular, [r] lacks a geminate counterpart in Dogri and is realised as a tap, a brief and rapid articulation. Additionally, the [rVm] and [mVn] sequences are voiced and homorganic/near-homorganic, creating a sonorant-to-nasal environment which is conducive to gestural overlap and articulatory blending. This overlap likely reflects a form of speech economy, whereby the C1 gesture is absorbed into the following nasal. Historically, Dogri has shown a tendency to assimilate certain homorganic CC, resulting in the formation of a geminate, e.g., OIA, [sərpɑ] to MIA, [səpp(ə)] (*serpent*) (cf. Table 1). Lastly, since the contrastive weight in such a word pair falls on the geminate [nn], which serves as the primary cue for lexical distinction, the C1 may be reduced in duration because of this phonological emphasis.

V2, though, consistently shows reductions among geminates, but the difference in contrast with the singleton is negligible (6 ms shorter for geminates on average). The minor shortening is attributed to the fact that geminates are longer by nature and consume more of the word-level time window. In order to maintain an overall prosodic timing in time-sensitive languages, the vowel is compressed to a certain degree. For all four phonemes, i.e. C1, V1, C2, and V2, mean durations in the form of violin plots by place of articulation for both word types are presented in Figure 4 a, b, c, and d, respectively.

Separate LMER tests were run to assess the effects of total word duration and voicing effects on C2 and V1 durations with random intercepts and slopes for Word Type by Speaker.

The test run to assess the effects of Word Type and Place on total word duration revealed that there were significant effects of Word Type, $F(1, 8.99) = 56$, $p < .001$ and Place, $F(3, 1542.82) = 7.26$, $p < .001$, on the total word duration. There was also a strong interaction effect, $F(3, 1539.51) = 3.08$, $p = .02$, indicating the duration to be longer for geminates than singleton at all articulation places with an average of 54 ms.

The observations about the total word durational difference between S–G pairs show that geminate words are longer despite contractions in the preceding vowel, as presented in Figure 5a. On average, geminates exceed singleton by 53 ms, with place-specific differences: Bilabials (41 ms), Dentals (55 ms), Retroflexes (53 ms), and Velars (66 ms). Longer geminate words imply that length distinction is part of the phonological grammar in Dogri in the environment when V1 is short, is actively maintained and not neutralised or redistributed in the mental lexicon of speakers, thus preserving the durational prominence of the lengthened consonant. It shows that geminates are not a variant of a single consonant but carry an additional prosodic weight to the word.

Since the stimuli included target words embedded in both Fixed and Meaningful Carrier Phrases, it was observed that both word types exhibited longer closure durations in the Fixed Phrase condition. However, with minor lengthening among singleton in the

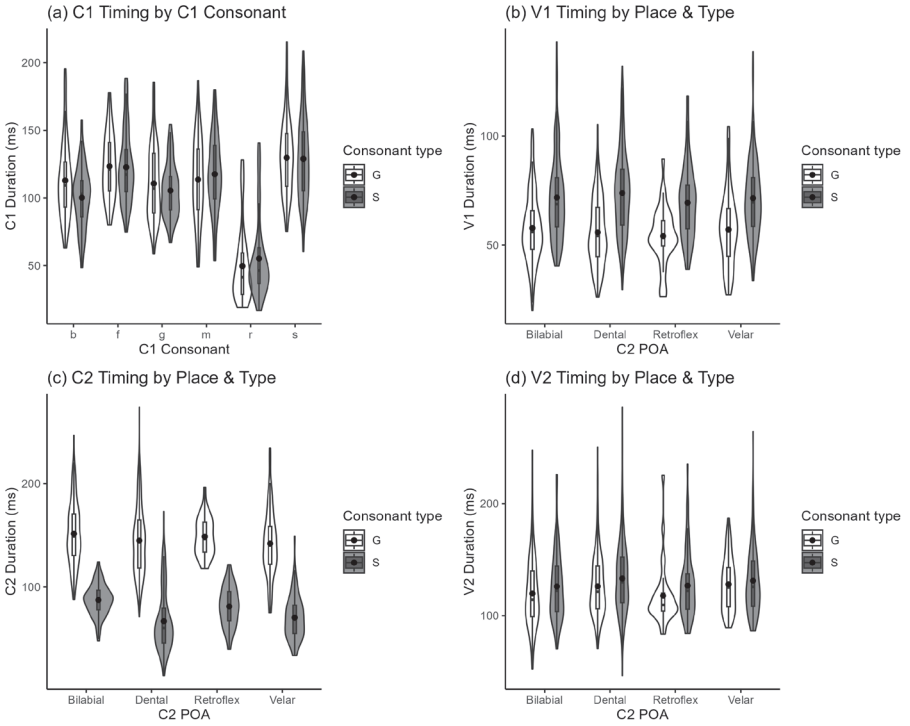


Figure 4 C1–V1–C2–V2 mean duration by articulatory place in singleton–geminate consonants.

Fixed Phrase context as opposed to when in the Carrier Phrase, the contrast is statistically insignificant (Bilabials: 2.45 ms, Dentals: 3.1 ms, Velars: 13.94 ms). Geminates, contrastively, show much longer duration; words tend to be significantly longer in Fixed Phrase (cf. Figure 5 b). This pattern was maintained across all places of articulation for C2 (Bilabials: 13 ms, Dentals: 16.4 ms) except for Velars: 9.1 ms. A plausible interpretation is that as opposed to singleton, geminates within fixed phrases are usually produced with greater articulatory care, likely due to the more controlled and isolated nature of the context, involving only two flanking words, i.e. ʊs ___ sʊneːjɑː, compared to the faster, natural, and more fluent speech characteristic of semantically coherent carrier phrases. The natural speech may overlap segments or facilitate coarticulation, leading to an overall shorter duration of both word types in carrier phrases. Thus, geminate consonants appear more sensitive to speech rate, emphasis, or prosodic phrasing than singletons.

The model was run to assess the effects of Voicing (voiced vs. voiceless), Word Type, and Place (Bilabial, Dental, Velar) on C2 and V1 durations. For C2, the analysis output significant main effects of Voicing, $F(1, 1421.28) = 808.4$, $p < .001$, Place, $F(2, 1421.31) = 28.95$, $p < .001$, and Word Type, $F(1, 9.23) = 169.13$, $p < .001$. In addition, significant interactions were observed: Voicing × Place, $F(2, 1421.41) = 89.41$, $p < .001$, Voicing × Word Type, $F(1, 1421.30) = 17.71$, $p < .001$, Place × Word Type, $F(2, 1421.30) = 23.35$, $p < .001$, Voicing × Place × Word Type, $F(2, 1421.40) = 9.30$, $p < .001$. These results indi-

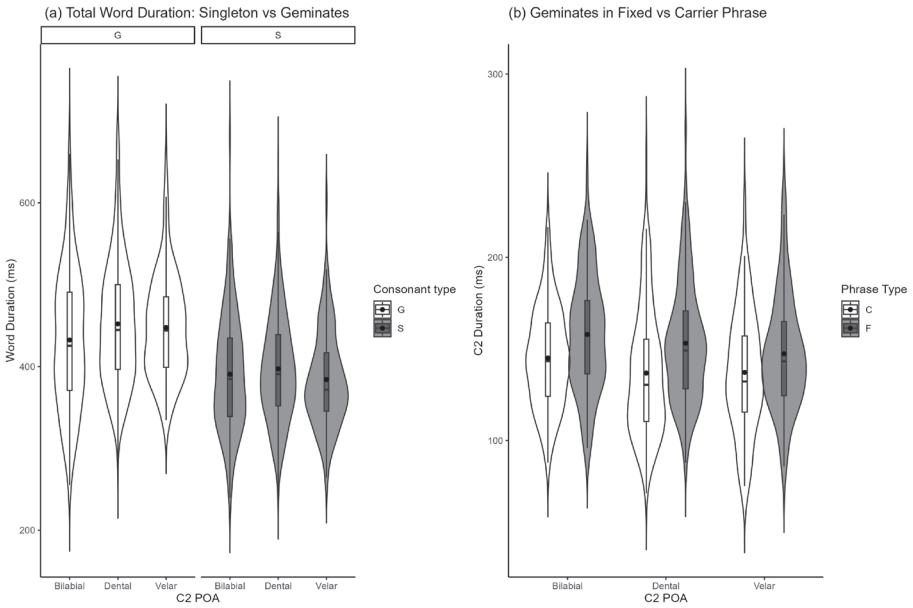


Figure 5 a) Total word duration (C1+V1+C2+V2) among Dogri singleton and geminates; b) Dogri geminates among fixed vs. carrier phrases.

cate that C2 duration is significantly influenced not only by voicing, place of articulation, and word type individually, but also by their interactions, suggesting that the voicing effect on C2 duration differs by both place and singleton/geminate status.

Concerning the effect of voicing on length-contrasting word pairs, it is revealed that C2 among voiceless singletons and geminates is consistently longer than the voiced counterparts in Dogri. While voiceless word-medial consonants among geminates are reported to be 36.8 ms longer than their voiced counterparts, whereas C2 among voiceless singleton maintained an average of 48 ms and 28 ms for Dentals and Velars, respectively, with minor lengthening of 5.8 ms among Bilabial singleton pair (statistically non-significant, $p = .66$). Among Bilabials, the minimal pair with C2 [p]–[m], despite their voicing contrast, shows minute lengthening effects for voiceless singleton may be attributed to the intrinsic durational properties of nasals like [m], which are generally shorter compared to voiceless stops like [p] due to the manner difference. In other words, a stop, in this case, is a burst type followed by aspiration, whereas a nasal is accompanied by continuous voicing, making the pair not ideal to show robust contrast on C2 duration. This theory is validated by the observation that, among voiced–voiceless geminates, bilabials show the least duration separation (23 ms) as opposed to Dentals (46.2 ms) and Velars (41.2 ms). Thus, both phonemes are qualitatively different, regardless of the word type (S/G).

Similarly, for effects of voicing on V1 duration, the analysis revealed significant main effects of Voicing, $F(1, 1421.10) = 291.02$, $p < .001$, Place, $F(2, 1421.11) = 9.87$, $p < .001$, and Word Type, $F(1, 9.31) = 20.88$, $p = .001$. A significant interaction was also found between Voicing and Place, $F(2, 1421.12) = 19.88$, $p < .001$. However, the interaction

between Voicing and Word Type was not significant, $F(1, 1421.12) = 2.84$, $p = .09$, while the Place × Word Type interaction, $F(2, 1421.09) = 1.85$, $p = .15$, and the three-way interaction (Voicing × Place × Word Type), $F(2, 1421.11) = 1.36$, $p = .2$, were also found to be non-significant.

The impact of voicing on the preceding vowel further attests that V1 is shorter before voiceless singleton and geminates, a pattern reflecting cross-linguistic tendencies of pre-consonantal vowel shortening before voiceless obstruents (see Shrotriya, 1995; Maddieson & Gandour, 1975 for Hindi; Ghosh, 2015 for Bengali). On average, V1 among singleton voiceless words are 14.5 ms shorter than the voiced counterpart, with a place-specific duration difference of 16 ms for Bilabials, 10 ms for Dentals, and 17.6 ms for Velars. V1 among voiceless geminates, on the other hand, are found to be 11.8 ms shorter on average. This finding attests that the vowel preceding the consonant tends to be shorter. The influence of C2 voicing on Consonantal (C2) and Vowel Durations (V1) is presented in Figure 6. Retroflex consonants were excluded due to the absence of a corresponding voiced retroflex word pair in the stimuli.

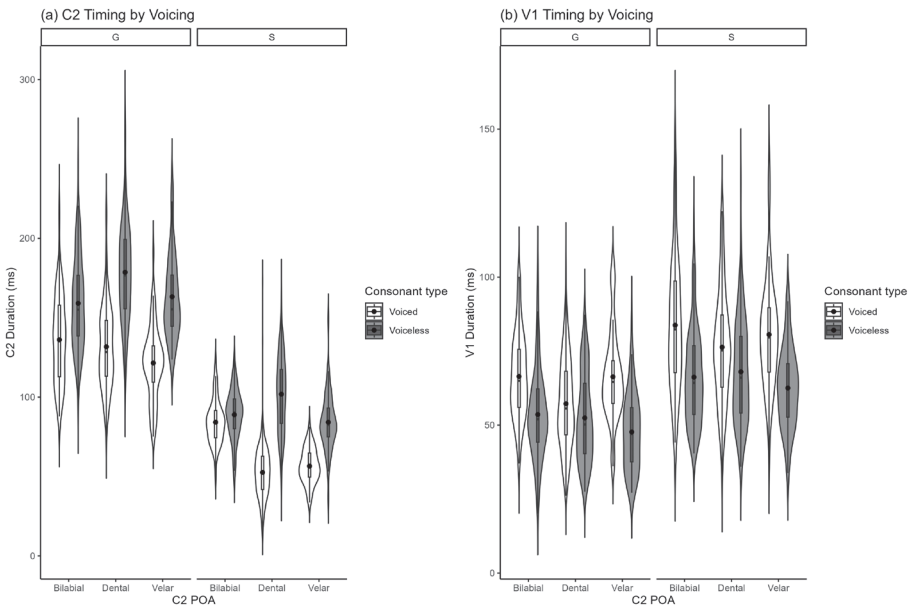


Figure 6 Effects of C2 voicing on consonantal (C2) and vowel (V1) durations.

4. Conclusion

In conclusion, the contrast between Dogri singletons and geminates is phonetically marked by geminates being almost twice as long as singletons, accompanied by pre-consonant vowel shortening. The duration of the word-initial consonant and the word-final vowel remains stable across both word types, indicating no anticipatory lengthening for

geminates for the former. The voicing contrast between the two word types aligns with cross-linguistic patterns; voiceless geminates exhibiting longer durations and the preceding vowels showing shortening compared to their voiced counterparts. Within Fixed vs Carrier Phrase contexts, geminates show longer duration for the former. The longer total word duration for geminates evidences that gemination is not merely a local (segmental/phonetic) phenomenon but rather has distinctive phonological weight. In other words, despite the non-compensatory nature of the initial consonant, shortening of the preceding vowel and of the final vowel among geminates, geminates are realised with an overall temporal increase, rather than "through" redistribution, reflecting their phonological status within the language.

REFERENCES

Aldubai, N. A. (2015). The impact of geminates on the duration of the preceding and following vowels in Ta'zi dialect. *Arab World English Journal*. Vol. 6.

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

Bhatia, T. K. (1993). *Punjabi: A cognitive descriptive grammar*. Routledge.

Blevins, J. (2004). *Evolutionary phonology: The emergence of sound patterns*. Cambridge University Press.

Blevins, J. (2005). The typology of geminate inventories: Historical explanations for recurrent sound patterns. In *Seoul Linguistics Forum 2005* (pp. 121–137). Language Education Institute, Seoul National University.

Boersma, P., & Weenink, D. (2023). Praat: doing phonetics by computer. Version 6.3.16, Retrieved from http://www.praat.org [Computer program]

Cho, H. (2016). Variation in vowel duration depending on voicing in American, British, and New Zealand English. *Korean Society of Speech Sciences*, *8*(3), 11–20.

Dulai, N. K., & Koul, O. N. (1980). *Punjabi Phonetic Reader*. Central Institute of Indian Languages.

Engstrand, O., & Krull, D. (2009). Durational Correlates of Quantity in Swedish, Finnish and Estonian: Cross-Language Evidence for a Theory of Adaptive Dispersion. *Phonetica*, *51*(1–3), 80–91.

Ghai, V. K. (1991). *Studies in Phonetics and Phonology with special reference to Dogri*. Ariana Publishing House.

Ghosh, A. (2015). Acoustic correlates of voicing and gemination in Bangla. In D. M. Sharma, R. Sangal, & E. Sherly (eds.), *Proceedings of the 12th International Conference on Natural Language Processing (ICON-2015)* (pp. 413–418). NLP Association of India. https://aclanthology.org/W15-5957

Gleason, H. A., & Gill, H. S. (1969). *A reference grammar of Punjabi* (3rd ed.). Punjabi University.

Grierson, G. A. (1904). On the modern Indo-Aryan alphabets of North-Western India. *Journal of the Royal Asiatic Society of Great Britain and Ireland*, *1* (67–73).

Gupta, V., & Chowdhary, S. (2024). *History of translation in Dogri literature: An overview*. In NTM, CIIL, Mysore (pp. 8–21).

Hamzah, M. H., Fletcher, J., & Hajek, J. (2016). Closure duration as an acoustic correlate of the word-initial singleton/geminate consonant contrast in Kelantan Malay. *Journal of Phonetics*, *58*, 135–151.

Han, M. S. (1994). Acoustic manifestations of mora timing in Japanese. *The Journal of the Acoustical Society of America*, *96*(1), 73–82.

Hirata, Y. (2007). Durational variability and invariance in Japanese stop quantity distinction: Roles of adjacent vowels (sokuon, or moraic obstruent). *Onsei Kenkyu (Journal of the Phonetic Society of Japan)*, *11*(1), 9–22.

Hussain, Q. (2015). Temporal characteristics of Punjabi word-medial singletons and geminates. *The Journal of the Acoustical Society of America*, *138*(4), EL388–EL392.

Idemaru, K., & Guion-Anderson, S. (2010). Relational Timing in the Production and Perception of Japanese Singleton and Geminate Stops. *Phonetica*, *67*(1–2), 25–46.
Kaur, K., & Dwivedi, A. V. (2018). Dogri and its dialects: A comparative study of Kandi and Pahari Dogri (Linguistics Edition, 115). LINCOM.
Kluender, K. R., Diehl, R. L., & Wright, B. A. (1988). Vowel-length differences before voiced and voiceless consonants: An Auditory Explanation. *Journal of Phonetics*, *16*(2), 153–169.
Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13), 1–26.
Lahiri, A., & Hankamer, J. (1988). The timing of geminate consonants. *Journal of Phonetics*, *16*(3), 327–338.
Lenth, R. (2025). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.11.1-00001, https://rvlenth.github.io/emmeans
Ladefoged, P., & Maddieson, I. (1996) *The sounds of the world's languages.* Oxford: Blackwell Publishers.
Maddieson, I., & Gandour, J. (1975). Vowel length before stops of contrasting series. *The Journal of the Acoustical Society of America*, *58*(S1), S61-S61.
Mitterer, H., (2018) The singleton-geminate distinction can be rate dependent: Evidence from Maltese, *Laboratory Phonology 9*(1).
Ohala, M. (2007). Experimental methods in the study of Hindi geminate consonants. In M.-J. Solé, P. Speeter Beddor, & M. Ohala (eds.), *Experimental approaches to phonology* (pp. 351–368). Oxford University Press.
Ridouane, R. (2007). Gemination in Tashlhiyt Berber: an acoustic and articulatory study. *Journal of the International Phonetic Association*, *37*(2), 119–142.
Ridouane, R. (2010). Geminates at the junction of phonetics and phonology. In C. Fougeron, B. Kühnert, M. D'Imperio, & N. Vallée (eds.), *Laboratory phonology 2010* (pp. 61–90). De Gruyter Mouton.
Rojczyk, A., & Porzuczek, A. (2019). Durational properties of Polish geminate consonants. *J. Acoust. Soc. Am.*, *146*(6), 4171–4182.
RStudio Team. (2023). *RStudio: Integrated development environment for R (Version 2023.09.0).* RStudio. https://www.rstudio.com
Samudravijaya, K. (2003). Durational characteristics of Hindi stop consonants. In *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech 2003)* (pp. 81–84).
Shrotriya, N., Siva Sarma, A. S., Verma, R., & Agrawal, S. S. (1995). Acoustic and perceptual characteristics of geminate Hindi stop consonants. In K. Elenius & P. Branderud (eds.), *Proceedings of ICPhS95* (vol. 4, pp. 132–135). Arne Strömbergs Grafiska.
Shrotriya, N., Verma, R., Gupta, S. K., & Agrawal, S. S. (1996). Durational characteristics of Hindi consonant clusters. *Proceeding of Fourth International Conference on Spoken Language Processing* (vol. 4, pp. 2427–2430).
Khattab, G., & Al-Tamimi, J. (2014). Geminate timing in Lebanese Arabic: the relationship between phonetic timing and phonological structure. *Laboratory Phonology*, *5*(2), 231–269.
Tabain, M., & Jukes, A. (2016). Makasar. *Journal of the International Phonetic Association*, *46*(1), 99–111.
Turco, G., & Braun, B. (2016). An acoustic study on non-local anticipatory effects of Italian length contrast. *The Journal of the Acoustical Society of America*, *140*(4).
Warikoo, K. (2000). Tribal Gujjars of Jammu & Kashmir. *Himalayan and Central Asian Studies*, *4*(1).

APPENDIX

Test word pairs with a short pregeminate vowel (CV̲CV:)

	Phoneme	Place	Manner	Voicing	Singleton	Geminate
1	p	Bilabial	Stop	Voiceless	təpɑ:	təppɑ:
2	p	Bilabial	Stop	Voiceless	tʃəpe:	tʃəppe:
3	t̪	Dental	Stop	Voiceless	kut̪e:	kut̪t̪e:
4	ʈ	Retroflex	Stop	Voiceless	pʰəʈi:	pʰəʈʈi:
5	t̪ʰ	Dental	Stop	Voiceless	kət̪ʰɑ:	kət̪ʰt̪ʰɑ:
6	k	Velar	Stop	Voiceless	səkɑ:	səkkɑ:
7	d̪	Dental	Stop	Voiced	səd̪ɑ:	səd̪d̪ɑ:
8	d̪	Dental	Stop	Voiced	gəd̪ɑ:	gəd̪d̪ɑ:
9	g	Velar	Stop	Voiced	bəgɑ:	bəggɑ:
10	l	Dental	Lateral	Voiced	kəlɑ:	kəllɑ:
11	m	Bilabial	Nasal	Voiced	rəmi:	rəmmi:
12	n	Dental	Nasal	Voiced	sʊne:	sʊnne:
13	n	Dental	Nasal	Voiced	məni:	mənni:

RESUMÉ

Tato práce zkoumá fonetické a fonologické vlastnosti kontrastů mezi jednoduchými a zdvojenými souhláskami (gemináty) v jazyce dogri v pozici, kde předchozí samohláska je krátká, tj. ve slovní struktuře CVCV:. Statistické výsledky ukazují, že zdvojené hlásky jsou časově téměř dvakrát delší než jejich jednoduché protějšky (~ 70 ms), přičemž dochází ke konzistentnímu zkrácení předcházející samohlásky (~ 16 ms). Zbývající segmenty, tj. souhláska na začátku slova a samohláska na konci slova, zůstávají u obou typů slov bez významných změn.

Další zjištění týkající se efektů znělosti, kontextu na úrovni fráze a celkového trvání slov u obou typů potvrzují obecné trendy znělosti: neznělé gemináty jsou delší s přidaným důkazem zkrácení samohlásky před souhláskou ve srovnání s jejich znělými protějšky. Navíc gemináty vykazují významně delší trvání v kontextech pevných frází než v kontextech nosných frází. A konečně, delší celkové trvání slov u geminát (~ 53 ms) naznačuje, že geminace v dogri přesahuje lokální fonetický jev a má výrazný prozodický a fonologický význam. Tyto výsledky podporují názor, že časové prodloužení odráží jejich status jako fonologicky kontrastivních jednotek, spíše než že by bylo vedlejším produktem přeskupení trvání segmentů.

Pranav Badyal
Institute of Phonetics
Faculty of Arts, Charles University
Prague, Czech Republic
pranav.badyal@ff.cuni.cz

2025 ACTA UNIVERSITATIS CAROLINAE PAG. 195–219
PHILOLOGICA 3

INTERVIEWS WITH BERND MÖBIUS AND ZDENA PALKOVÁ ON LIFE IN PHONETICS

PAVEL ŠTURM, JÜRGEN TROUVAIN

ABSTRACT
This article features two in-depth interviews with long-time phoneticians, Bernd Möbius and Zdena Palková, who reflect on their careers and the changes they have witnessed in the field. Although their professional paths differ widely – from speech technology to speech on the stage, among others – both scholars share a strong connection to the core methods and questions of phonetics. The interviews touch on the development of the phonetic sciences over several decades, discussing shifts in research focus, academic culture, and international collaboration. By taking an open-ended, conversational approach, the interviews reveal not only the interviewees' scientific insights but also personal experiences, motivations, and views on the future of the discipline. The article highlights the value of interviews as a complementary format for documenting the history and diversity of phonetic research.

Keywords: interviews; history of phonetics; oral history

1. Introduction

Although interviews with phoneticians and other linguists occasionally appear in the literature (such as the conversations with William Labov, Peter Ladefoged, or Michael Ashby, see Gordon, 2006; Kaye, 2006; Ashby & McElvenny, 2022), they remain a surprisingly rare form of scientific publication in phonetics. However, we believe that interviews can be a powerful means of scientific communication.

First of all, they give space for experienced researchers to look back over several decades of work, often providing a retrospective understanding of how the field has evolved. At the same time, these conversations often turn towards the future: what still needs to be done, and what directions seem promising. For readers, interviews also create a chance to compare their own experience and viewpoints with those of the interviewee, especially when it comes to how we structure research or make decisions about what is important in our work. Furthermore, unlike standard research papers, interviews can provide space for broader conceptual discussions, differing views on methodology or theoretical frameworks, and personal perspectives on the field's development. Finally, they bring in the personal side of academic life: how careers unfolded, what certain moments felt like, and how relationships with mentors, colleagues, or institutions shaped someone's path.

https://doi.org/10.14712/24646830.2025.28

The authors previously developed the written interview project *Phonetic Sciences in Retrospect* (Trouvain & Šturm, 2025), which explores the evolution of the field through first-hand accounts from experienced scholars. The aim was to build a broader historical and conceptual understanding of phonetic sciences by gathering responses to a structured questionnaire organized around ten thematic areas. Using the same structure for each interviewee ensured consistency and allowed for direct comparison across responses.

In contrast, the current paper takes the form of oral interviews, which allow for a more interactive and dynamic exchange. This format makes it possible to explore specific points in greater depth and follow up on unexpected or particularly interesting insights during the conversation. While the written interviews focused more on overarching issues in the field, the oral interviews presented here offer richer glimpses into the interviewees' professional paths, working environments, and personal experiences – privileging small, individual insights over broad generalizations. We hope they provide both inspiration and new perspective on what it means to spend a life in phonetics.

In this paper, we present interviews with two prominent phoneticians, Bernd Möbius (BM) and Zdena Palková (ZP), each looking back on a long and varied career in the field. BM and ZP come from different backgrounds – in terms of age, gender, country, and research interests. Still, there are also many parallels. Their stories highlight the rich diversity within the phonetic sciences, while also revealing shared concerns and points of connection across the field.

2. Method

2.1 Participants

Bernd Möbius (BM) is a German phonetician born in 1959 in Andernach, West Germany. From 1979 to 1985, he studied Communication Research and Phonetics (major), with minors in Linguistics and Sociology, at the University of Bonn, where he also completed his PhD in 1992. Further academic stations include Bell Laboratories (New Jersey, USA) and University of Stuttgart (Germany), before being appointed full professor of Phonetics at Saarland University in Saarbrücken (Germany) in 2011. Since 2025, he has been holding the position of senior professor in Saarbrücken.

Zdena Palková (ZP) is a Czech phonetician born in 1938 in Olomouc, Czechoslovakia. From 1956 to 1961, she studied Czech and German at Charles University in Prague. Her entire academic career has been closely tied to the Institute of Phonetics in Prague, where she held various academic roles and served as the director of the institute from 1999 to 2007. In 2015, she was granted the title of Professor Emerita of Phonetics.

2.2 Procedure

Both interviews were held over two days in June 2025. Each conversation lasted approximately 60 to 90 minutes. The interview with BM took place online via a video chat (MS Teams) and was conducted in English. BM participated from his home in Germany, while both interviewers, JT and PŠ, joined from the Institute of Phonetics in Prague. The

session was audio- and video-recorded, and an orthographic transcript was automatically generated using Copilot (Microsoft, 2025), visible only to the interviewers during the conversation. This initial transcript provided a useful foundation for preparing the final written version.

The following day, the interview with ZP took place in person at the Institute of Phonetics in Prague. It was conducted in Czech by a single interviewer (PŠ). The conversation was audio-recorded using MS Teams, and once again, an automatic transcription was generated via Copilot to support the preparation of the final text.

The post-processing of the interviews involved several steps:

1. Editing the raw transcript to create a readable version by removing repeated words, correcting proper names, and inserting appropriate punctuation to reflect natural syntax.
2. Verifying the transcript against the audio/video recordings, with particular attention to unclear or uncertain passages.
3. Polishing the text by removing redundant content and reducing the number of discourse markers (e.g. *so, but, I mean*) to improve readability while preserving the speaker's style.
4. Formatting the interview for publication, including clear differentiation between the interviewee's and interviewers' contributions and appropriate paragraphing.
5. Final review by the interviewee, who was given the opportunity to make last adjustments or refinements to their responses.
6. Translation of the Czech interview into English by PŠ, in consultation with ZP to ensure accuracy and clarity.

3. Results and discussion

3.1 Common features

BM and ZP come from different cultural and institutional backgrounds and have followed distinct professional paths. Yet, several important commonalities emerge in their careers, reflecting shared experiences, values, and research themes (see Table 1).

Table 1 Common features in the careers of Bernd Möbius and Zdena Palková.

Aspect	Bernd Möbius (BM)	Zdena Palková (ZP)
Institutional commitment	Long-term roles at Bonn, Bell Labs, Saarland University	Lifelong career at the Institute of Phonetics, Prague
Focus on prosody	Research on intonation models, prosodic features in technology	Studied prosodic phrasing and rhythm in Czech
Applied phonetics	Applied phonetics to speech synthesis and speech technology	Applied phonetics in public speaking and theatre
International engagement	Active at ICPhS and Interspeech since 1991	Participated in nine ICPhS congresses since 1967
Technological adaptation	From diphone synthesis to neural networks	From manual analysis to digital tools in low-resource conditions

First, while BM held positions at multiple institutions, and ZP spent her entire career at the Institute of Phonetics in Prague, both demonstrated strong, long-term commitment to their academic homes. They were involved in teaching, research, and academic leadership, including serving as department heads. Closely related is their influence on the next generation of phoneticians. BM describes working with PhD students as one of the most enjoyable and rewarding aspects of his career, emphasizing mutual learning. Similarly, ZP highlights the formative role of her early teaching experiences, especially working with foreign students of Czech phonetics.

Second, both share a sustained interest in prosody, particularly rhythm and intonation, as a central element of spoken language. They approached this topic from different angles but with complementary aims. BM worked on intonation models and their integration into speech technology (e.g., adapting Fujisaki's model for German). ZP also focused on the suprasegmental level of Czech, especially in terms of prosodic phrasing and rhythm in text interpretation. Her overarching concern has been the relationship between units of spoken language (such as stress groups and prosodic phrases) and units of written text.

Equally significant is their commitment to the practical application of phonetic knowledge beyond academia. BM contributed to the development of German text-to-speech synthesis at Bell Labs and has continued to apply phonetics in speech synthesis and recognition. ZP has worked for decades with actors and broadcasters to improve spoken performance, notably serving as a phonetic consultant at the Czech National Theatre since 1990 and collaborating on over 200 productions.

Another shared feature is their strong international engagement, reflecting a commitment to staying connected to the global phonetics community. BM has regularly participated in conferences such as ICPhS and Interspeech, and served as editor-in-chief of *Speech Communication*. ZP attended nine ICPhS congresses, starting with the 1967 Prague meeting, and remained internationally active even under difficult political and economic conditions. As a side note, it's worth mentioning a subtle connection between Bonn and Prague: not directly through BM and ZP, but through an earlier visit by Milan Romportl – another phonetician from Prague – to the University of Bonn.

Finally, each experienced and adapted to major technological changes in the field, whether it was the rise of neural networks or the shift from analogue to digital tools.

3.2 Differences

There are several notable differences between the careers of BM and ZP – differences shaped by time, place, or academic context. Table 2 summarizes the key contrasts in their professional lives. However, it must be emphasized that some of these differences may also reflect the fact that BM and ZP belong to different academic generations, with nearly two decades between them.

One major difference lies in the academic environment of their respective countries. In Czechoslovakia, phonetics was an exceptionally rare field: it could be studied as a programme only in Prague, and since 1954 only as part of broader programmes such as Czech or German. In contrast, West Germany offered a dozen or so institutions where phonetics could be pursued as a dedicated subject, allowing BM to major in phonetics at the University of Bonn within a more structured academic setting. Today, thanks in

Table 2 Key differences in the careers of Bernd Möbius and Zdena Palková.

Aspect	Bernd Möbius (BM)	Zdena Palková (ZP)
Country and academic system	West Germany (later unified Germany), USA	Czechoslovakia (later Czech Republic)
Access to phonetics education	Studied phonetics in Bonn, one of ~10 institutions in West Germany	In Czechoslovakia, phonetics could only be studied in Prague
Career mobility	Held positions in multiple institutions (Bonn, Bell Labs, Stuttgart, Saarbrücken)	Spent entire career at the Institute of Phonetics in Prague
Research focus	Speech technology, intonation modelling, synthesis systems	Rhythm and intonation, applied phonetics (e.g., speech on the stage)
Use of technology	Worked with cutting-edge tools in well-funded environments	Worked with limited resources
International opportunities	Benefited from early access to international exchange	Faced travel restrictions, financial obstacles

part to ZP's post-1989 efforts, the Institute of Phonetics in Prague offers dedicated degree programmes in Phonetics at the Bachelor, Master and PhD levels.

Their career trajectories also reflect these institutional differences. BM held positions at multiple institutions in Germany and abroad, including a formative period at Bell Labs in the United States, before becoming a professor at Saarland University. ZP, by contrast, spent her entire professional life at the Institute of Phonetics in Prague, where she maintained continuity and professional standards through decades of political transformation and institutional change.

Their research interests likewise diverged. BM's focus was rooted in speech technology, including intonation modelling and speech synthesis. ZP focused on the practical description of spoken Czech, especially its prosodic aspects. She authored a book on the phonetics and phonology of Czech (Palková, 1994). Nonetheless, she also contributed to technological developments by providing phonetic feedback for early Czech speech synthesis systems, bridging academic phonetics and engineering practice.

These contrasting research paths also reflect broader disparities in technological and institutional resources. BM had access to advanced tools and well-funded labs, while ZP worked with limited equipment and often relied on her team's ingenuity – typical of the Czechoslovak academic setting at the time. A related difference lies in international engagement: BM benefited from institutional support and academic mobility, whereas ZP had to overcome political restrictions and often financed her own participation in conferences. Her attendance at nine ICPhS congresses stands as a clear testament to her determination to remain part of the global phonetics community despite the obstacles.

Finally, their intellectual influences reflect different scholarly traditions. BM was shaped by communication theory, linguistics, and speech engineering, often in interdisciplinary contexts. ZP's background includes linguistics and literary theory, which informed her phonetic perspective in distinctive ways. She views phonology – the study of meaningful sound distinctions – as an essential and natural complement to the articulatory, acoustic, and perceptual dimensions of phonetics.

In sum, the careers of BM and ZP highlight how personal interests intersect with broader cultural, institutional, and political conditions, resulting in two distinct yet equally valuable contributions to the field of phonetics.

4. Conclusion

The development of any scientific field – here phonetics – is always reflected by the development of individual careers and propelled by individual researchers. It is their ideas, interests and fresh perspectives that drive the field forward. The personal reflections shared by BM and ZP in these interviews offer a rare opportunity to look back on several decades of scientific activity in phonetics. As contemporaneous witnesses, their accounts provide a kind of evidence not typically found in reports of empirical studies and probably also not in overview articles. The interviews not only allow us to compare the two individuals, BM and ZP, but also changes in the field over time. In our opinion, comparisons between long-time overviews can serve as a valuable source of insight and can give us important impulses for future work.

Acknowledgements

This publication was supported by the Cooperatio programme of Charles University, Linguistics research area, implemented at the Faculty of Arts of Charles University, and also by the programme 'Ostpartnerschaften' funded by DAAD and organized by Saarland University. The authors would like to express their sincere gratitude to both interviewees for the time, insight, and care they contributed to this project.

REFERENCES

Ashby, M., & McElvenny, J. (2002). The emergence of phonetics as a field. In J. McElvenny (ed.), *Interviews in the history of linguistics: Volume I* (pp. 51–60). Language Science Press. https://doi.org/10.5281/zenodo.7092391

Barnes, J., & Shattuck-Hufnagel, S. (eds.) (2022). *Prosodic theory and practice*. MIT Press.

Daneš, F. (1957). *Intonace a věta ve spisovné češtině [Intonation and the Sentence in Standard Czech]*. ČSAV.

Gordon, M. J. (2006). Interview with William Labov. *Journal of English Linguistics*, *34*(4), 332–351.

Hála, B. (1967). *Výslovnost spisovné češtiny I. [Pronunciation of standard Czech I.]* (2nd ed.). Academia.

Hála, B., & Sovák, M. (1962). *Hlas – řeč – sluch [Voice – speech – hearing]* (4th ed.). SPN.

Janota, P. (1967). *Personal characteristics of speech*. Academia.

Kaye, A. S. (2006). An interview with Peter Ladefoged. *Journal of the International Phonetic Association*, *36*(2), 137–144. https://doi.org/10.1017/S0025100306002519

Microsoft (2025). *Microsoft 365 Copilot in Teams* [Computer software]. https://m365.cloud.microsoft

Palková, Z. (1994). *Fonetika a fonologie češtiny [Phonetics and phonology of Czech]*. Karolinum.

Romportl, M. (1978). *Výslovnost spisovné češtiny II. [Pronunciation of standard Czech II.]*. Academia.

Trouvain, J., & Šturm, P. (2025). 'Phonetic Sciences in Retrospect' – A written interview project. In J. Cęcelewski et al. (eds.), *HSCR 2025: Proceedings of the Seventh International Workshop on the History of Speech Communication Research* (pp. 27–34). TUDpress.

APPENDIX

A: Interview with Bernd Möbius

Bernd, you are now a freshly retired professor. Can you say a little bit about the beginnings of your career as a scientist, when you started studying speech, and what your interests were? And a few words about your different academic stations, please.

In retrospect, I guess, it was certainly not planned this way from the beginning, but a recurring theme of my research is integration of phonetic knowledge in speech technology. That's something I have always stated on my homepage as well.

Although that sounds a bit like a unidirectional influence, I mean it in both directions. I always wanted to advance speech technology based on fundamental research in speech science, but, conversely, also to use the technology to test our hypotheses of how humans process language and speech by means of tools that implement these hypotheses. So, for me, speech synthesis used to be the prime example of this kind of approach.

And when did you start studying the big topic of phonetic knowledge for speech technology?

Somehow this was also almost right from the beginning of my studies of phonetics and a few other things a long time ago in Bonn [the capital of West Germany at the time]. I started my studies in 1979.

So it was during the Cold War, right?

Exactly. I had just finished my military service in Germany. But perhaps I should start even a little bit earlier, because how did I actually come to this somewhat obscure field of study of phonetics?

Yes, please do.

Because I guess phonetics could be studied at perhaps ten places at most in Germany.

With Germany, you mean West Germany?

Oh, that's a very good point, yes, indeed. So perhaps it was even less than ten. And there were also a few places that were called something like 'Sprechwissenschaft,' which translates into 'speech science,' right? But with a lot of focus on spoken interaction, on conversation, and less in the technical sense, I think, that we have in our focus very often today.

But you know, I was not exactly exposed to phonetics. At school I was always interested in languages and grammar, and one day, while my best friend and I were doing our homework, the older brother of my best friend – let's call him Hans – was also doing homework, but for the university, and he was working with all kinds of weird symbols he put on paper by pen. I asked him what these strange and funny symbols were. He explained to me that these were symbols that help us transcribe spoken language for all languages in the world, basically in a uniform way, and he called them phonetic symbols. Of course, I had come across some of them in, say, English and French lessons at school, but never in a systematic way.

I see. And was the interest for those strange symbols so strong that you then decided: I can study that, let's do that? Or was there interest in some other aspects which motivated you to start the studies?

Well, I did ask him, of course, about the background, why he's using those symbols, and what it is that he's actually studying at the university. He told me more about that and

he was very helpful, eventually, because after my military service I had to make a serious decision about what to study. I was always torn between astronomy on the one hand, which was another sort of hobby horse for me as an adolescent, and something with language on the other.

I then went for languages, and Hans advised me to study phonetics as a major and linguistics as a minor, not the other way around. He himself, he did his bachelor (in today's terms) in Bonn and his master in Munich. So when we talked about this, he was already in Munich, as a PhD student with Hans Tillmann. Ironically, perhaps, the Bonn Phonetics Institute was housed in the former Astronomy and Observatory building. So, in a sense – how do they say it in English? – 'I had my cake and ate it.'

Were there other surprises when you were starting your studies of phonetics in the ancient observatory in Bonn?

I guess lots of surprises because everything was essentially new to me in phonetics, apart from those symbols with which I had familiarized myself a little bit already after I first came across them.

I think my first lecturer in phonetics, Dieter Stock, started his first session on acoustic phonetics with mathematical formulae that meant to describe the acoustic structure of a spoken utterance as a variation of whatever over time. He jumped right into something that, didactically, might have been better taught a bit later, making it easier for young students to approach the field – but I survived.

It was counteracted by my professor of phonetics, Gerold Ungeheuer, who didn't actually teach phonetics courses at this time; he had, from his point of view, advanced beyond that. He advocated phonetics as part of a larger area of research that he called 'communication research.' He approached that from both a technical and a social perspective. From him I learned that phonetics and linguistics are disciplines that are also couched in a much larger context, all of which has to do with how humans communicate by language.

In Bonn, you did your Magister, then also your PhD. How did you come to your PhD topic, and what was it, of course?

My PhD topic was to develop an adaptation of Hiroya Fujisaki's intonation model to the German language. A lot of factors contributed to that. Of course, when you need funding for a PhD project, you have to be a little bit opportunistic and that was both on my side and on the side of my PhD advisor, Wolfgang Hess, who applied for funding from the German Research Foundation (DFG) for a project with exactly this topic, and it was perhaps a hot topic at the time. You know, Fujisaki's model was very influential but also controversial, and so it was relatively straightforward to apply for funding for such a project and get it. So it was less me defining my topic for myself, on my own or in interaction with my advisor, and more like the advisor seeing this opportunity, introducing me to this model.

And were you happy with the topic? How did it go?

Oh yeah, yeah, absolutely. It was intonation and my Magister thesis was already on German intonation, but in a much less technical sense. It was more experimental – I was trying to follow the approach that Klaus Kohler in Kiel advocated for a short time, which was to find out what the phonological atoms of intonation are. He called them 'tones,' not exactly in the sense of ToBI [Tone and Break Indices]. It was basically at the same time

when Janet Pierrehumbert developed her phonetics and phonology of American English intonation, a famous PhD thesis. So Kohler had a different approach that was, I think, inspired by Halliday's work.

My idea was to run perceptual experiments to find out whether the posited tones have any reality in human perception and processing. That was more or less successful, and it was a bit critical of the overall approach. I don't think this is why Kohler eventually dropped this kind of approach for himself. It may have contributed to it, but I never talked to him about this actually, and I don't think my experiments could have a lot of influence on Kohler's thinking.

You said that during that time modelling intonation was a hot topic. How do you feel the temperature of the topic nowadays?

Nowadays, in a sense, it has completely disappeared. There was a collection of papers a few years ago in a book entitled *Prosodic Theory and Practice* (Barnes & Shattuck-Hufnagel, 2022). The chapters basically summarised the different models that had been proposed over many years and their current state. I don't think any of them really are actively used in research these days. Except, of course, ToBI and the ToBI-kind of approach is highly influential still and used by so many people for so many languages and successfully so. But at the time when I was doing my PhD, it was a real controversy and there was a real competition between models. I don't see that anymore.

In recent years, you do see many papers exploring the tonal and intonational structure of many languages and varieties, which you can find a lot at the phonetics congresses, for instance. But if you look at the leading journals like *Speech Communication, Computer Speech and Language*, perhaps even *Journal of Phonetics*, etc., there's this strong trend in very recent years (last two or three years, I would say) to find out how intonation or intonational features are represented in neural models. This seems to be the really hot topic these days.

It's not modelling in the sense of, you know, a model that predicts and generates intonation contours for applications like speech synthesis or as a component of a speech recogniser anymore. We know, more or less, how segmental acoustic-phonetic features are represented in these neural models. A logical next step would then be to ask how intonation is represented and whether it plays a role for the neural models in the end. In the layers that are close to the acoustics, to the input, you can see these features, but they tend to disappear and, hopefully, are merged into higher level, more abstract features along the way. But there doesn't seem to be a need to model intonational features or the intonation contour separately. It's very different.

What did you do after your PhD?

After my PhD, I stayed in Bonn for another year as a postdoc. They had just started a very huge broad-scale project on speech-to-speech translation, which became known as 'Verbmobil,' in which basically all labs doing something like phonetics, many linguistics labs and all the big industrial players in Germany were involved, like 50 partners from academia and industry. I worked there for a year but then I got an offer from Bell Labs in Murray Hill, New Jersey, USA.

That was in which year, Bernd?

This was in 1993. Wolfgang Hess was contacted by a colleague, Juergen Schroeter from Bell Labs, who asked if Wolfgang knew a young researcher who would be interested to

work on German speech synthesis at Bell Labs. And Wolfgang in turn asked me whether I would be interested, and without knowing what I would be getting into, I said, well, yes, sure, I'm interested but I do not know so much about synthesis. We had a synthesis project in Bonn working on diphone and demi-syllable synthesis, but I was only tangentially involved – Thomas Portele and Karl-Heinz Stöber were the researchers in this project. It served as a test bed for my intonation model, though.

Maybe I should make a very quick excursion to the very early days of my studies, because the guy that I mentioned, Dieter Stock, the lecturer in phonetics, was also very technical and he had a Votrax chip. That was a chip that implemented something like, maybe I'm wrong, formant synthesis and was able to speak German with a heavy American English accent. So there was an early exposure to something, to very early speech synthesis when I was a young student, but it was not a continuous exposure over my studies. During my PhD I was then exposed to diphone synthesis a little.

I found the offer very interesting, but I didn't really know what I would be getting into. It was a big step, of course, moving house, the family to the US. It was meant to be for one year, then they said, you know, 'German synthesis, it's a lot of work to do indeed, at least two years.' Actually, after half a year, they offered me a permanent position and I took it. Although it was not planned that way, it was in the end five years and three months or so before I decided to go back to Germany (for several reasons).

And were you happy with the output of the German synthesis system you'd built there?

At the time? Yeah, I think I was enthusiastic about it. I thought it was fantastic. Whenever I listened to it, I thought, how did I do that? It sounds so great. When I played it to my wife, the enthusiasm was a bit less, but she was exposed to it occasionally, so she was also not completely neutral. When I played it to anybody else not involved in speech synthesis, they said: 'Terrible! Okay, it's quite intelligible, but the quality is awful. I would never use such a system.'

But that was not the reason to return to Germany, I guess.

No, it was not. The time at Bell Labs was fantastic. I learned so much in relatively short time, unbelievable! But for family reasons I wanted to move back to Germany. Also, I had never planned to stay in the States forever, and even at that time I couldn't actually envision it.

I got an offer from Greg Dogil, the professor of phonetics in Stuttgart, whom I met several times during conferences, and at ICSLP (International Conference on Spoken Language Processing) in 1998, he said: 'There is an opening for an assistant professor in my group starting January 1999. It may be the unique opportunity for you if you really want to go back and pursue your career in academia, not in the industry. I would be delighted if you took it.'

I had to think about it. It was a hard decision for me to leave Bell Labs because I had friends there, not only very good colleagues. But my wife and I decided to move back to Germany and of course I don't regret it.

Was it hard for you to move from Rhineland (where Bonn is located) to the Swabian area (with Stuttgart as its centre)?

Perhaps harder than to the US! No, it wasn't. It wasn't really. It was a lot of fun, really. Especially the dialect differences were a constant topic at the lunch table, not only between Rhineland and Swabia – there were people from all over Germany and outside

of Germany, of course, in the lab. The cognitive map that people have, for instance of Germany, is sometimes very distorted. So I was often addressed as a 'Fischkopf' [the head of a fish], which is what people sometimes call those who live on the North Sea coast. Now, Rhineland is basically roughly in the middle in the north-south dimension in Germany, so pretty far from 'Fischköpfe'. But everything that's north of Mannheim or Frankfurt is probably close to the sea for Stuttgart people.

But you returned to Bonn, is that correct, for a short period after your Stuttgart years?

I obtained my habilitation quite early in Stuttgart after moving there. Habilitation is like the 'second book' that used to be required if you wanted to become a professor in Germany. I was something like an associate professor, but not permanent, not tenured in Stuttgart. I applied for tenured professorships in Germany and a few other places as well, including the position in Bonn. This had a lot of appeal because I would have come back to where I started as a student, also located in Rhineland.

So I accepted the offer, and I stayed there for three years as a substitute professor because I was never formally appointed. It was all hanging in the air for various legal reasons. However, I was absolutely lucky, in a sense – it was almost a coincidence. When it became clear that, in Bonn, they would never fill the position again and would destroy the institute (not only phonetics, also computational linguistics), I was asked, exactly at that time, by people in Saarbrücken whether I would consider applying for the professorship there – which I did.

You've mentioned a few names so far. Who would you say is your main mentor or your main mentors during your career time?

I cannot single out one person. I really owe so much to several people, perhaps even many people. I already mentioned my first phonetics lecturer, Dieter Stock. He not only introduced me to acoustic phonetics, but he also offered me a desk in the lab, so I could stay in the lab all day. He offered me research assistant jobs, unpaid, but still fantastic. So, I assisted in acoustic phonetic analysis of German dialect data for *Mittelrheinischer Sprachatlas*, then for somebody else's habilitation thesis that I think never materialised, and in sonographic analysis of bird vocalisations using a Kay Sonagraph. And I got access to a PDP 15 computer that was also instrumental, an early exposure to computers as tools for phonetic analysis. When I said I got access to PDP 15, I meant it literally: it had a door that you could open and step into the computer – it was a room-sized computer, an amazing beast.

Dieter Stock also programmed his own early version of something that we later got to know as 'ESPS X-waves' from 'Entropics' or even later as 'Praat' that everybody uses now. He programmed it first in assembler, later in Fortran, and that's how I got a little bit into programming. So, he was very influential in lots of senses, I guess. I have already mentioned the Votrax chip that could synthesize German with an American accent. He was also my de facto advisor for my Magister thesis.

I also mentioned my first phonetics professor, Gerold Ungeheuer, saying that his lectures were more about communication theory rather than phonetics proper. He was also not accessible to beginner students. Although in principle he had something like office hours, you would have to make an appointment many weeks in advance and then he wasn't actually there. He passed away before I started my third year as a student. Despite this, he was still very influential because he provided me with this broader view of the field, the larger perspective of speech communication.

My PhD advisor was Wolfgang Hess and I already talked about him. Of course, he was very, very influential on me and very supportive. He encouraged me to take the offer from Bell Labs. And this is where the influence became more than individual, because I was an integral part of the speech synthesis group, comprising very well-known people like Jan van Santen (I think he was my primary mentor when I started my job there), Richard Sproat, Julia Hirschberg, Chilin Shih... I should stop mentioning names because I will then have to leave out a few others who would also deserve being mentioned. I definitely should mention one more name, that's Joe Olive who was the head of the group and who actually brought me there and was also always very supportive, even though he was disappointed, I think, when I decided to go back to Germany. But he understood perfectly.

Of course, this was an environment where you could just walk down the hall to talk to famous people and ask them for advice regarding some problem you had run into. They were always friendly and helpful. It was a fantastic environment.

You were influenced, of course, by others, but you were probably also influential yourself, for example on PhD students. Maybe it is not always pure fun to work with PhD students and reading their PhD theses. But eventually, in my experience, as supervisors we always learn something from them. Do you have an example for that?

You're absolutely right and also wrong. You were right in that I definitely learned from the work of my PhD students, and I hope they learned a bit from me too. However, you were also wrong because you suggested that it is not always fun to read these theses. I assume you mean reading chapters over and over again or something like that, not necessarily the final product. But I never thought that this is not fun, that it's just a duty that I'm doing; on the contrary, I usually enjoyed it. Still do.

For me, it is a bit difficult sometimes to find the right balance between interfering, perhaps too much, especially on earlier versions of chapters, and leaving too long a leash. And then having the problem of capturing the people again when they got lost a little in the various interests that they have and lost focus a bit.

Supervising PhD students and working with them is, maybe, the top enjoyable part of our profession overall. I think there are many other enjoyable aspects, but this is where I think I may have had the most lasting influence on other people. The interactions with all my PhD students were productive, friendly and, overall, entirely enjoyable.

Some were a bit more remote from my own interests. But then I became interested in these topics even more. Of course, you know, as PhD students proceed and make progress, they become the real specialists in their area, much more so than the advisor. So the advisor learns from the students.

You are also a rather active editor, for instance of the journal Speech Communication, and guest editor of, for example, special issues in journals. What is your view on how the scientific output should nowadays be published in the optimal way? Books, journals, or conferences? And should an article be available on arXiv or something similar or should it always be published in a peer-reviewed way?

You know, I'm getting older and almost necessarily a little bit more conservative. I tend to think that many aspects or components of scientific publishing that I've come to learn about or got to know over the many years have survived the storms of time. They have proved to be very valid and effective publication channels. But some things have changed, of course.

I think the importance of books has diminished. Of course it depends on the type of book, but even textbooks for students are less popular these days than they used to be when I was still an early-career researcher. I have a certain dislike for collections. Depending on the scope and topic, they still have their place, definitely. But I also have had a number of rather negative experiences as a contributor to such collections – some of them never materialised, and all this work, writing chapters, was almost for nothing.

The top journals in our field, I think, still play – and have to play – a very crucial role: for peer-reviewed scientific publishing, for quality, for making sure that the scientific quality is top. Peer reviewing – yeah, definitely, we need this system. Although we know it has drawbacks, especially from the view of an editor, I can't see a better system than that.

I see the tendency to upload pre-publication versions of papers on arXiv or similar platforms from a negative perspective, I must say. It seems to be taking on its own life in a sense. Other researchers rely a lot on these pre-publication versions, taking them for gospel even though they have never been peer-reviewed, never been checked, never been approved by authorities (the peers). Of course, I also see why it can be useful. For instance, if you really have a groundbreaking contribution to make, you want to stake the claim, you want to make sure that you can claim ownership of this original idea, then putting the stuff on arXiv is one way of achieving that. But I've also seen many arXiv papers that have never been published for whatever reason or that have never been updated after being published. I see more downsides than upsides in this development.

What about conferences? Can you say a few words about them? Perhaps even some repeated shortcomings that you identify in the presentations or in the organization?

I really can't say anything negative about it. I think these conferences, especially the big ones in our field like the phonetics congresses and Interspeech, still have a very important role to play and I think they achieve these goals very well. I think most of the presentations (both oral and poster) are very professional, very good, very well prepared. Naturally, there are always negative outliers, but in general the vast majority is in my experience very good.

The conferences are so big now, the ones that I mentioned, that, obviously, you have to be very selective to build your own schedule, and you will necessarily miss a lot of papers that you would also have found interesting but you could not have started an interactive discussion with the authors. You can read the papers, of course, in the proceedings afterwards.

Regarding organization, I think a recurring problem is poster sessions. These are often very tiresome, the acoustic conditions are often very bad. It's difficult to organize these poster sessions in an ideal way, but at some conferences it's better, at others it's problematic.

Over time, I've come to appreciate small venues more and more, and also the value of such small conferences and workshops – especially workshops where you can also present unfinished and not so polished work. You have a lot of interaction with the other participants, who are usually specialists in exactly this research area, and you get feedback from your colleagues while you're still working on these topics. This can be very fruitful.

So they are often presentations without published proceedings, right? Perhaps just extended abstracts.

Very often, yes. I would always encourage PhD students to attend these kinds of workshops that specialize in exactly the areas that they work on. Of course, I also see the need

for them to publish in peer-reviewed conferences and hopefully, at the later stages of their PhD, in journals as well. They have to find the right balance. They should not only look at the opportunity to publish, but also to get into stronger interactions with more advanced researchers on the one hand and their peers, other PhD students, on the other. That's also very valuable at the smaller events.

Let's return to speech synthesis. Nowadays, synthetic speech sounds quite different compared to the early 90s. Would you say that the work on formant synthesis, diphone synthesis, maybe also on unit selection was paving the way for the synthesis nowadays? And would you still see that phonetic knowledge can contribute to improvement of speech synthesis quality nowadays?

That's very complex. To be a little bit facetious, I think that the early approaches to synthesis with articulatory synthesis, concatenative synthesis, statistical parametric synthesis, etc., don't have a direct impact anymore on current synthesis techniques. The knowledge that we gained from these earlier approaches was invaluable, was immense, but I don't think it translates directly into the current synthesis approaches. Rather, it was invaluable for the researchers who work on speech science to understand how humans process speech.

If you look at the researchers and the authors and the research teams and their affiliations in speech synthesis papers these days, at the mainstream, this not the same cohort of people as before. They have brought a lot of progress that enables us to produce synthesized speech that is sometimes indistinguishable from natural speech, a progress that 10 years ago I would not have believed to be possible in such a short time. It came from a completely different angle.

Coming back to what I said at the very beginning, a speech synthesizer – in the old-fashioned way a modular system – can be a fantastic vehicle for testing your hypothesis or local theory about certain components of human speech processing. If you build an implemented model of one component in human speech processing or production, say intonation, and you put that as a module into the speech synthesizer, and if it improves the synthesizer, making it sound much more natural in terms of prosody, then you must have understood something about what humans do. You must have done something right in the overall system. And in this way, you can go from module to module.

I think the best example perhaps is articulatory synthesis, which is also the most ambitious one. You need to have so many partial models of the human speech production process, each of which is a major challenge. And if you do something wrong, it will probably percolate through the system; conversely if you improve your system in terms of output quality, then it's almost certain that you understood something correctly and implemented it correctly.

I think that's a very nice example and also a very nice topic for ending our conversation. Thank you very much, Bernd, for sharing your insights and for being our conversational guest today.

B: Interview with Zdena Palková

To begin with, a few questions about your field. You've been active at the Institute of Phonetics in Prague since 1961, and you're still involved today.

Now as an emeritus professor, which means I'm only here occasionally.

Could you share with us how you actually got into phonetics?

I entered the Faculty of Arts (back then it was called the 'Philological Faculty') in 1956. The system of study had already changed earlier, after 1949; in philology, only combinations of languages were available. All study programmes were double-subject and always pedagogically oriented. So in addition to the main subjects, lectures in pedagogy and psychology were included – essentially preparation for gaining a teaching qualification. It wasn't possible to pursue academic studies at all. My combination was Czech and German.

Personally, I was mainly interested in Czech and literary theory. That's actually what I originally applied for, based on an older post-war lecture catalogue. But by then, the programme no longer existed. It was a subject that had fascinated me already at grammar school. What really attracted me to the Faculty of Arts was literature – not language.

When mentioning post-war lecture catalogues, phonetics was listed there, wasn't it?

Yes, it must have been, but I wasn't looking for it. When I started at the faculty, I had no idea what phonetics even was. But I do know that the last graduate who studied phonetics as a subject was the orientalist Petr Zima. After 1954, phonetics was no longer offered as a separate subject. Our double-subject language studies were unevenly weighted: one of the subjects was the major, with the final thesis. In my case, that was Czech studies. And within that, my interest was always directed towards literature.

I first encountered phonetics right at the beginning of my studies, in the first semester. Professor Hála gave the lectures – very engaging and easy to follow. After I passed the phonetics exam, I thought to myself, 'Great, that's done, I understand it – and I'll probably never deal with it again.' I had the feeling that everything in phonetics had already been figured out. Professor Hála's lectures for beginners left no room for uncertainty. Later, when I started lecturing myself, I made a point of highlighting those uncertainties instead.

That there's always more to discover.

Yes. Then, in the fourth year, it was time to start thinking about the diploma thesis. Students either received a topic or could propose one themselves. At that time, I was interested in the structure of prose texts from a rhythmic perspective. I had the idea that analyzing the spoken interpretation of a text – performed by several speakers, especially non-professionals – might reveal something interesting. So I proposed a topic focused on trying to characterize rhythmic devices in Vladislav Vančura's novel *Markéta Lazarová*, because I felt his text was particularly well suited for that purpose.

But the Department of Czech Literature wasn't willing to accept such a topic, arguing that it was too interdisciplinary. However, the phonetician Milan Romportl was open to it. And that's how I ended up at the Phonetics Institute, which at that time was still part of the Department of Czech Language. There's a saying: 'Once the claw is caught, the whole bird is trapped.'

When you started working on your thesis with Milan Romportl, did that also mean you began attending some phonetics courses?

No, because there weren't any. The study programme only included the basic introductory courses – elective seminars in the field came later. But I did start working at the Phonetics Institute as a student research assistant. I was really taken with the insti-

tute's active and welcoming atmosphere. And suddenly, I had the opportunity to record and play back texts for my thesis. A portable tabletop tape recorder was a rare piece of equipment at the time.

Did you consult your thesis work with Professor Romportl? Did he give you any training?

At the time, Milan Romportl had an exceptional amount of his own work, including a major time-sensitive publication. He was always willing to answer questions if I had any, but there simply wasn't time for systematic supervision. When it came to the things I immediately needed for the experimental part of my thesis, I was kindly and helpfully guided by Přemysl Janota. It's worth noting that, back then, university students were expected to be much more independent than they are today. I had more than enough work just processing the material and searching for and reading the necessary literature – and on top of that, I still had to finish my studies. In the fourth year alone, we had thirty-two mandatory hours per week; only the fifth year was a bit more flexible.

My thesis was well received, and after I graduated, Professor Bohuslav Hála offered me a position as an assistant. I was lucky that a position had just opened up at the institute, as there was otherwise a hiring freeze across the faculty. It wasn't until I started working there that I could really begin to systematically build up my knowledge of the field. I remember how useful my first major task was, assigned to me by Professor Hála: to compile a subject index for the book *Voice, Speech, Hearing*, which was just coming out in its fourth edition (Hála & Sovák, 1962).

When did you start teaching phonetics?

I didn't teach during the first three years. That was actually a wise rule at the time, and in my case, a necessity. Assistants weren't supposed to teach for this period, and the institute adhered to that. When I did start, I was first assigned the basic course in Czech phonetics for foreign students, which turned out to be very useful for me. Having to explain things really made me aware of many connections. And hearing the effects of cross-language interference in practical exercises helped me better understand some of the specific features of Czech, which might otherwise have remained just theoretical claims.

When you were starting out at the Institute, who inspired your work?

Given the way my interest in the field developed, I drew inspiration from distinguished figures across different areas of philological scholarship – some even during my student years – mostly within Czech studies. Chronologically:

An excellent foundation in linguistic study was provided to us by Vladimír Šmilauer, a Czech studies scholar and onomastician, but for me primarily a syntactician. Through the way he guided us in syntactic analysis, he taught us to understand the methodological steps involved in an objective, independent analysis of language in a text. Thanks to that, I later came to understand the importance of vagueness in language description. Concepts, categories, and descriptive relationships are defined precisely – but we also have to account for the possible occurrence of vagueness in how they're applied to actual linguistic material. And the question may arise: what level of vagueness is still acceptable, so that we don't need to redefine the description of a given phenomenon?

Another exceptional figure in terms of my early development was Jan Mukařovský, a scholar of verse and aesthetics, with a methodological focus on structuralism. The official ideology during my studies rejected the structuralist approach and, with it, Mukařovský's most important works. But for me, his *Chapters from Czech Poetics* were

a very instructive read. I found his concept of metrical impulse particularly inspiring for Czech verse structure analysis. I was delighted to discover that Professor Mukařovský was lecturing in the Czech studies programme, and I enrolled in his elective seminar, even though I knew it wasn't theoretical but historical in focus. When I later gave a report in the seminar on my thesis topic, Professor Mukařovský not only wasn't put off by its interdisciplinary nature – he actually gave me several helpful suggestions.

When I later became a permanent member of the Institute of Phonetics, Přemysl Janota became an invaluable mentor. I owe him a great deal for gradually and critically introducing me to the field in all its breadth and variety, as well as for familiarizing me with the principles of experimental work. Přemysl Janota himself was a versatile phonetician and also a speech therapist, with a primary interest in speech sound analysis, auditory perception, and individual timbre. He designed several devices that made listening easier; most notably the highly valued *speech segmenter*. This device made it possible to move a listening window – or alternatively, a muted segment – along a loop of magnetic tape, with both the window length and the shifting method adjustable.

From a slightly later period, when I started to get a better sense of the broader landscape of linguistics at the time, especially Czech linguistics, I'd like to mention another figure who, from my perspective, was very important. The linguist František Daneš, a bohemist, syntactician, and text linguist, was also an intonation specialist. His study *Intonation and the Sentence in Standard Czech* (Daneš, 1957) helped strengthen the foundations for suprasegmental research in Czech phonetics. But what I found most instructive was the way he engaged in academic discussions. He was a passionate debater, often in disagreement with others. But he always truly listened to the other person's argument and took their counterpoints into account. That was rare.

What topics interested you over the course of your career, beyond your teaching duties?

In my own research, I focused mainly on the suprasegmental level of description, especially within the sound structure of Czech: defining and organizing units of analysis (such as the stress group, prosodic phrase in two levels, or completed utterance), examining how these sound units relate to one another, and how they connect with linguistic units in the text. In terms of methodology, given the limited technical resources at the time, perceptual tests were the most practical option. But in reality, there was relatively little time left for research.

Another branch of my professional work grew out of practical spoken language use, again based on Czech, this time covering both the segmental and suprasegmental levels. One path involved taking a closer look at the features of Czech as a foreign language, in a way that could be applied in teaching – both in terms of description and in preparing learning materials. The other path, which I felt more drawn to personally, was the phonetic side of language culture in contemporary Czech: both through descriptive work and practical involvement in public media, and later in speech on the stage.

Were these, for example, people working in radio?

The Institute had a tradition in that regard. It was established by Professor Hála right after the war. Radio announcers would come to the Institute for pronunciation training. I didn't witness that period myself, but recordings from those sessions have been preserved in the archive, and I later used them in lectures. The most important achievement in this area was undoubtedly Hála's effort to stabilize the codification of Czech pronunci-

ation, which eventually led to the publication of the *The Pronunciation of Standard Czech* (Hála, 1967). Milan Romportl then continued this work, focusing on the pronunciation of foreign words (Romportl, 1978).

Almost everyone took part in the practical activities related to speech culture – often courses and lectures for various public institutions. Bohuslav Hála, for instance, regularly conducted courses for funeral speakers, while Milan Romportl focused on wedding speakers. The Institute had recently acquired a portable Tesla Sonet Duo tape recorder. I remember how I used to search through individual speeches, following Prof. Hála's instructions, to find segments with mistakes, which he would then analyze and I would play back for the seminar participants. At that time, the possibility of on-the-spot speech recording was a novelty that drew considerable attention from the audience. The tradition of promoting speech culture continued until the 1990s. I myself, for example, led regular seminars at the Radio for nearly fifteen years for staff from several editorial teams.

What range of activities did the Institute cover back then?

The effort to practically support good spoken Czech was just one part – definitely not the central focus – of the Institute's work. I think Professor Hála, in his role as director, very deliberately aimed to make sure that all the key areas of phonetic description were being developed within Czech phonetics; at that time, mainly physiological phonetics and speech acoustics. With the same determination, he made sure that knowledge about Czech was kept up to date with international developments in the field, both in cross-linguistic comparisons and in the description of both the segmental and suprasegmental levels.

Bohuslav Hála himself was a firm supporter of so-called experimental (instrumental) phonetics, in the tradition of Josef Chlumský. As for the new approaches to phonological description of sound structure that had strong roots in the interwar Prague structuralist school, let's just say he was sceptical. But when Milan Romportl often leaned in that direction, Hála accepted it as a natural thing. Looking back now, I think the parallel development of both major approaches to describing the sound level of spoken language – instrumental phonetics and phonology – as I encountered it when I entered the field, was one of the fortunate strengths of Czech linguistics as a whole. The fact that their followers occasionally got into heated arguments wasn't crucial. Neither camp was dominant enough to suppress the other, and when necessary, they were able to exchange useful information.

Abroad, it is common for researchers to move between institutions. Can you comment on your experience of having worked at the same institution for half a century?

In a word: You develop a sense of responsibility for that institution.

What was the atmosphere like at the Institute when you first arrived?

When I came to the Phonetics Institute, the impact of the major changes that had taken place – and were still taking place – in higher education and scientific institutions was already noticeable. We had, of course, been aware of them as students.

The first thing that struck me at the Institute was the sense of a strong and deliberately maintained tradition: an effort to uphold traditional standards in both the quality of work and workplace relationships. Prague phonetics was well known and respected abroad; it used to be a very well-equipped institution, with a reliable level of research and teaching.

Due to external changes, it lost its degree programme and its administrative independence; it could no longer decide its own future in terms of equipment, tasks, and to a large extent even its research orientation.

The people working there did everything they could to minimize the negative consequences for the field. What struck me most was the positive working atmosphere, the supportive environment, and the determination to take advantage of every new opportunity that arose for the benefit of the discipline. This sometimes took on a humorous side, like when we spent afternoons untangling high-quality studio magnetic tapes discarded by the Radio and 'spinning' them back into usable reels for our studio tape recorders.

If we jump ahead ten or twenty years, how did the working conditions change?

The institute had a stable staff of 5–6 people, and until the late 1980s we didn't get any additional positions – I remained the youngest until I was fifty. Such a situation is literally fatal for the field. It was similar with the equipment. The last modern innovation had been two portable tabletop tape recorders at the beginning of my career. Then, only around 1988–89, we acquired one of the first simple desktop computers. Not because the school cared about the phonetics discipline, but simply because there was still no interest in it at the faculty then.

A fortunate fact for Prague phonetics was that Přemysl Janota was also a skilled engineer and could adapt available devices for phonetic work. For example, his segmenter – a very clever device that made repeated listening easier – gained international recognition.

Which instruments commonly used abroad were clearly missing?

We didn't have a sonagraph. We were probably the only phonetics institute in Europe, if not in the world, without one.[1] The closest sonagraph occasionally available to us was in Prešov, Slovakia. In Prague, there was one at the phoniatric clinic, but using it required complicated requests and long waiting times. It was easier to simply go, for example, to Dresden.

What did you replace it with, if at all?

You can always calculate the spectrum manually using harmonic analysis based on Fourier. But I'm joking. Still, much can be calculated, for example from oscillographic recordings. Přemysl Janota wrote several papers relying on acoustic analysis, usually designing original innovations from the available equipment. For his publication *Personal Characteristics of Speech* (Janota, 1967), he built a simple synthesizer that allowed him to manually set and simulate any individual vowel composed of five tones, based on specified frequencies and with controlled duration and loudness.

A truly game-changing moment in the development of fields dealing with sound was the digitization of the audio signal and its computer processing. This brought a major shift in what was possible, especially for a place like our institute – small, fairly underfunded, and not favoured by institutional authorities. Still, we already had some experience with computer speech processing since we had long worked with acousticians studying the speech signal. I remember visiting the big mainframe computer at the Research Institute

[1] JT & PŠ: At the Institute of Phonetics in Saarbrücken, there was also no sonagraph. First spectrographic analyses were performed with digital devices at the beginning of the 1990s.

for Communication Technology several times, where the first attempts at Czech speech synthesis were taking place.

But the new possibilities that came along later were substantially different. The advanced demands for describing phonetic phenomena in all kinds of acoustic processing – especially handling the suprasegmental layer of continuous texts – required specialized and costly equipment. When using computers, though, more accessible specialized software could do the job. Plus, computers with rapidly increasing memory made it possible to process large datasets. For the research activities of the Phonetics Institute, this basically meant a lifeline. By a lucky coincidence, these changes happened almost simultaneously with a turning point in the social and political scene in our country.

Were there other changes at the Institute in the 1970s and 1980s besides the technological lag?

Fortunately, the working atmosphere within the Institute remained unchanged. However, there were two significant and sad changes in personnel: in 1970, Bohuslav Hála passed away, and already in 1982, so did Milan Romportl. What unfortunately did begin to change over time was the overall atmosphere at the faculty. When I first started, being part of the academic community at the Faculty of Arts, Charles University, was regarded as a prestigious responsibility. Many outstanding figures worked across various disciplines. Students came with respect and expected demanding work. A natural consequence of this was a certain collegial decency that extended across the academic community as a whole, including administrative staff. I think this helped to some extent buffer the restrictive pressures coming from above. As the older generation of academic personalities gradually disappeared, this sense of responsibility visibly weakened, although this varied across departments and disciplines.

How did the fact that the phonetics department was part of a larger unit affect its functioning?

I would say that in terms of its academic orientation, it didn't have any significant impact. Phonetics is – whether that's an advantage or a drawback – a fairly self-contained field by its very nature. Whether collaboration arose or not didn't depend on administrative affiliation. At first, the Institute was included as part of the Department of Czech Language. Later, comparative and general linguistics, as well as Czech for foreigners, were added. This overly large unit was eventually split again, and for a long time the Institute functioned as part of the Department of Linguistics and Phonetics. But collaboration between linguists and phoneticians wasn't any better or worse in either structure.

However, the coexistence of multiple disciplines within a single unit also has a second, very material dimension: staff positions, teaching loads, funding. And there, obstacles certainly existed.

If I may, let's move on to phonetic congresses. By coincidence, your first one was held in Prague, in 1967. What do you remember most vividly about it?

Once again, it was that strong sense of a binding tradition. Officially, the congress was held under the auspices of the Czechoslovak Academy of Sciences, but the main organizers were university people: Bohuslav Hála served as President, Milan Romportl as Secretary General, and Přemysl Janota as Secretary. Most of the actual work was done by the Phonetics Institute, and there was a clear, shared effort to ensure that the congress lived up to the pre-war reputation of the Institute.

There were no concerns about the academic level of the event. The programme was rich, clearly structured into thematic sections, and balanced as much as possible in terms of the languages covered. Representatives from more than thirty countries attended. The worries came from elsewhere – namely, the economic and political situation in the country. The economic conditions at the time were, to put it mildly, poor. The gap between everyday life in Prague and in major cities of Western Europe was already quite visible. Throughout the congress, we did our best to keep that gap from showing, at least during the official programme. When choosing venues, we relied heavily on the beauty of the city.

There were also more serious concerns, especially for Romportl and Janota as the main organizers. It was important to allow experts from the West to attend, possibly even Czech émigrés. The state security service was given a preliminary list of all congress participants and granted permission. It was 1967, a time of relative thaw in international relations – otherwise the congress couldn't have happened at all – but there was still a lingering fear of some kind of politically sensitive incident. You could feel the tension in the Institute right up until the end of the event. Luckily, nothing happened.

How about the other congresses? You attended a total of nine.

The congresses were very important to us not only from a professional point of view. Even in the most difficult times, the Institute made a point of staying visible within the international phonetics community. Attending congresses, held regularly every four years, offered a good opportunity to do so. We always tried to present at least one paper – preferably more – and to arrange for as many people as possible to attend in person, even if only passively.

After the Prague congress came Montreal (1971), where Romportl was the only one from the Institute. I didn't attend in person until Copenhagen (1979). That time, we travelled as a group – it was called 'professional tourism,' or something like that. It was a joint trip for which participants were granted an exit permit because it had a work-related purpose. Such trips usually included people from several different institutions and had to be recommended and approved by the Ministry of Education.

At the next six congresses, I presented a paper of my own: Utrecht (1983), Aix-en-Provence (1991), Stockholm (1995), San Francisco (1999), and Barcelona (2003).

So you were covering the costs yourselves? Or did the Institute pay?

The Institute didn't have its own funds – any contribution had to come from the Ministry via the faculty. Sometimes they contributed, but based on their own (non-professional) criteria. Most of the time, we paid the expenses ourselves. But at least we were allowed to travel.

Starting with the congress in Aix-en-Provence (1991), we were finally able to attend all the following ones freely, on our own passports. Later on, we even received some financial support from the university, more or less, depending on what grants the Institute managed to secure. My last congress was Saarbrücken (2007), where we attended a meeting within the international project Sound to Sense (S2S), a Marie Curie Research Training Network, in which the Institute was involved at the time. And of course, in 2023, I followed the 20th congress back in Prague – this time just as a guest.

Which congress did you like the most?

I think the best-organized congress I experienced was Copenhagen. It was led by Professor Eli Fischer-Jørgensen, and phonology was a hotly debated topic. The number of

participants and the range of thematic areas still felt manageable, 'human' – rich, but not overwhelming. It was possible to get an overview and choose what interested you. Surely, this was thanks to the great effort of the organizers: carefully scheduling presentations so that related topics wouldn't overlap in time. For example, if the sessions on linear and nonlinear phonology are held simultaneously, while their essence is to be discussed in relation to each other, the interested attendee won't be thrilled. Such cases multiplied when specialized firms took over organizing the ever-expanding congresses and controlled not only the logistics but also the programme. Maybe it started in Sweden? I'm not sure. But I was probably least happy with the organization in Aix-en-Provence – it was the hardest for me to follow the flow of events there.

Which congress did you like best in terms of social atmosphere, meaning the participants and the informal parts of the programme?

The organizers of all the congresses I attended made a real effort to ensure that participants felt welcome, had a good time outside of the sessions, and saw as much as possible of the host city. And all the places I visited as part of this 'professional tourism' had something remarkable to offer: Barcelona, San Francisco, all of them.

But congresses also had another memorable aspect. Many colleagues returned time and again. We knew each other. It became part of our professional curiosity to see who turned up at the next congress – and who didn't. And to ask why, if they were missing. As I've already said, attending congresses gave our institute a chance to maintain international contacts. That proved to be incredibly valuable later on.

How do you see the change in topics at the congresses?

I think it's a result of how much technical possibilities for professional and interdisciplinary communication have evolved. Probably an inevitable consequence. At earlier congresses, it was fairly easy to define clear and distinct thematic areas, and you could also tell which issues were central at the time and which were more on the margins. As the congresses grew larger, the content of the sessions started to scatter more and more. Just take a look at the long list of sections – within one congress, the same topic might appear in multiple places, under different titles, with only minor shifts in focus. That lack of clarity was already noticeable in San Francisco. Back when we didn't have easy access to foreign academic literature, congress proceedings were a valuable way to get a solid overview of current topics across the broad field of what's called the phonetic sciences. That kind of clarity would be impossible now.

The very concept of what a congress is seems to have changed, too. It used to be that research teams wanted to present what they considered their most important results – sort of their calling card. When I watched the 2023 Prague congress, my impression was that quantity outweighed quality. It felt less like showcasing a lab or spotlighting a problem, and more like an opportunity for the younger generation to practice giving presentations. And to see a bit of the world.

Substantive debate, presentation of results, and exchange of ideas now seem to happen online. That's not a flaw – it's simply a change. Probably a natural one. I think the same applies elsewhere today, not just in the phonetic sciences.

Which foreign institutions did you maintain contact with?

I'll leave out the Slavic studies contacts, which had a long tradition and mostly meant visits to us, not trips abroad. Thanks to Milan Romportl, there was ongoing cooperation

between the Institute and several institutions in East Germany, especially in Halle/Saale. Contacts in the western direction were more of a continuation from earlier times and were tied to individual people.

What form did that cooperation take?

The institute in Halle/Saale focused broadly on spoken language ('Sprecherziehung') and on cultivating standard German pronunciation. Access to East Germany was relatively easy. The institute in Halle regularly organized conferences, which we often attended. I myself spent a two-week internship there already before the Prague congress. Conversely, colleagues from the German Democratic Republic (i.e., East Germany) were happy to come to Prague. Through Halle, we also had contacts with Jena. There was further cooperation with East Berlin and with the Technical University in Dresden. Milan Romportl himself also spent a year as a visiting scholar in Bonn (West Germany).

Professor Hála had numerous professional ties with France, and later these were maintained by Marie Dohalská, a specialist in French. She spent time in France on multiple occasions, including a long-term research stay. The English specialist Alena Skaličková often recalled her meeting with Daniel Jones and kept in touch mainly with institutions in London. Přemysl Janota spent part of his studies in the Netherlands, with Professor Louise Kaiser. Even though later he wasn't allowed to travel for many years, he preserved warm connections with Dutch colleagues even from afar (Nijmegen, Amsterdam, Utrecht). He also taught Dutch phonetics and grammar at the faculty. His other connections led to Sweden (to Professor Fant's team) and to Norway. One figure who must be mentioned in this context is Professor Martin Kloster-Jensen, a Norwegian phonetician from Bergen (who also worked in Bonn, Hamburg, and Oslo). He was very fond of the Institute and of Prague, and somehow always found a way to visit us, no matter the circumstances.

In 1970, Professor Antonie Cohen offered me a one-year research stay in Utrecht. However, the ministry found out that my sister had emigrated, and that was the end of it.

How did the Institute find its footing after the fall of the regime?

When the socio-political situation changed dramatically, we were happy, but we couldn't ignore the fact that, after all those years, the Institute was on the verge of professional collapse, and people completely exhausted. But it turned out that our efforts to keep Prague phonetics visible internationally hadn't been in vain.

Soon after, friends reached out with practical support. Professor Johan Liljencrants came to visit from Sweden and generously gave us access to his state-of-the-art acoustic spectrum analysis software, free of charge for research and teaching. We had a computer, but no money to buy such programmes ourselves. Then we heard from Professor Hans-Walter Wodarz in Frankfurt am Main, an émigré and graduate of Prague phonetics. In a short time, he secured substantial funding to help us buy new academic literature. Personal connections started turning into institutional ones.

Starting in the 1990s, I had several productive research stays in Frankfurt am Main, and over time, also gave a number of talks in Jena and to the Slavic scholars in Bonn.

You also apply phonetics in theatrical practice. Could you tell us more about your collaboration with the National Theatre?

It's not only the (large) National Theatre, but also smaller stages. I started in a very intimate setting, at the Theatre on the Balustrade around the 1979/80 season. I've been working at the National Theatre continuously since 1990. Working with actors in smaller

theatres is a bit different from working at the big stage, though not as drastically as people sometimes assume. My phonetic collaboration isn't what's commonly referred to as 'voice coaching' – actors bring that with them from school – but primarily it's about working with the text. I provide actors with something like feedback on their spoken performance, including discussions about possible corrections or the implications of using a different variant.

A basic responsibility is to make sure the actors speak clearly enough to be easily understood. Unfortunately, this task is more relevant today than it used to be. Everyday spoken Czech is often not easily intelligible anymore. It has sped up, mostly at the cost of very sloppy pronunciation. The ambition among actors to 'speak well' is also gradually declining. At the beginning of my work (at least at the National Theatre), it was almost a given: 'If I speak on stage, I must be understood even in the second gallery.' Actors themselves used to request feedback. Some of today's graduates from acting schools (often stars of TV series) not only lack this ambition, but they believe that clear pronunciation wouldn't sound natural – according to the current buzzword, 'authentic.'

Czech theatre doesn't really have a tradition of speech consultants, as they (hopefully still) do in Germany. I consider it a professional success that this supporting role is now seen as a regular part of production at the National Theatre.

Are you present at all rehearsals?

Not all of them, but I'm there quite often, depending on how rehearsals are going (and what my schedule allows). At certain stages, I try to follow rehearsals continuously. For example, during the early table-read sessions when the text is being analyzed. And later on, once longer sections of the play are being run continuously on stage and the rehearsal isn't being stopped all the time. Naturally, I want to be at all the so-called main and general rehearsals. And I never skip the public general rehearsal with an audience. – I might skip some of the blocking rehearsals, where the actors are mostly figuring out their positions on stage and often just improvise the text (which they may not even know by heart yet).

Do you focus on anything other than intelligibility?

I absolutely have to make sure that the pronunciation matches the style of the production. For instance, the epenthetic /v/[2] can't be allowed in a Shakespearean text, unless it's a line from a character who, even in the original, is linguistically marked as a folk or lower-class figure. In a classical play costumed in the style of Mary Stuart's court, traces of a Prague accent or, say, a Moravian dialect would be completely out of place. But even when a particular dialect forms the basis of the spoken text, it's important to monitor the degree and manner of its use. Speech on the stage isn't a copy of reality; it's a functional stylization.

But what I focus on the most – and what also requires the most effort, because it's often the most challenging for the actors themselves – is the relationship between the written text of the play and its spoken realization on stage; the shifts in meaning that happen along the way between those two poles. Actors are quite receptive to comments in this area; in fact, it's often the best ones who ask for them. They know that the way they perceive themselves may not be how the audience interprets their lines. A phonetician who

[2] A non-standard process in Czech; e.g., /okno/ (*window*) realized as [vokno] rather than [ʔokno].

attends the rehearsals can point out problematic spots in the text before an actor settles on a version that would later need to be changed. I think that's the main reason the actors don't see my presence as a nuisance, but as genuinely helpful.

And that relationship to the text brings us to a lovely conclusion. Thank you for such an inspiring conversation.

RESUMÉ

Článek představuje dva rozsáhlé rozhovory s význačnými fonetiky Berndem Möbiem a Zdenou Palkovou, kteří se v nich ohlížejí za svou profesní dráhou i proměnami oboru, jichž byli svědky. Jejich odborné cesty se sice výrazně liší – od výzkumu v oblasti řečových technologií až po zkoumání řeči na jevišti –, oba však sdílejí pevné ukotvení v základních metodách a otázkách fonetiky. Rozhovory reflektují vývoj fonetických věd v průběhu několika desetiletí, proměny výzkumných priorit, akademické kultury i mezinárodní spolupráce. Otevřená, dialogická forma přináší nejen odborné postřehy obou osobností, ale také jejich osobní zkušenosti, motivace a představy o budoucnosti disciplíny. Článek zároveň ukazuje, že rozhovor je cenným formátem pro dokumentaci dějin a rozmanitosti fonetického výzkumu.

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

Jürgen Trouvain
Language Science and Technology
Saarland University
Saarbrücken, Germany
trouvain@lst.uni-saarland.de

ACTA UNIVERSITATIS CAROLINAE
PHILOLOGICA 3/2025

Edited by Radek Skarnitzl, Jan Volín
Cover and layout by Kateřina Řezáčová
Published by Charles University
Karolinum Press, Ovocný trh 560/5, 116 36 Prague 1
www.karolinum.cz, journals@karolinum.cz
Prague 2025
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 38 Prague 1, Czech Republic
(books@ff.cuni.cz)