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 pr̝̊ 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 presents 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.

<table>
<tr><th>Phone</th><th>Observed duration (ms)</th><th colspan="2">Inherent duration (ms)</th><th>Local articulation rate</th></tr>
<tr><td>k</td><td>72</td><td rowspan="2">136</td><td></td><td>0.98</td></tr>
<tr><td>o</td><td>67</td><td rowspan="2">109</td><td>1.03</td></tr>
<tr><td>n</td><td>33</td><td></td><td>1.09</td></tr>
<tr><td>t͡ʃ</td><td>110</td><td colspan="2">114</td><td>1.04</td></tr>
<tr><td>iː</td><td>80</td><td colspan="2">68</td><td>0.85</td></tr>
</table>

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 $LAR = dur_{inh} - dur_{obs.}$

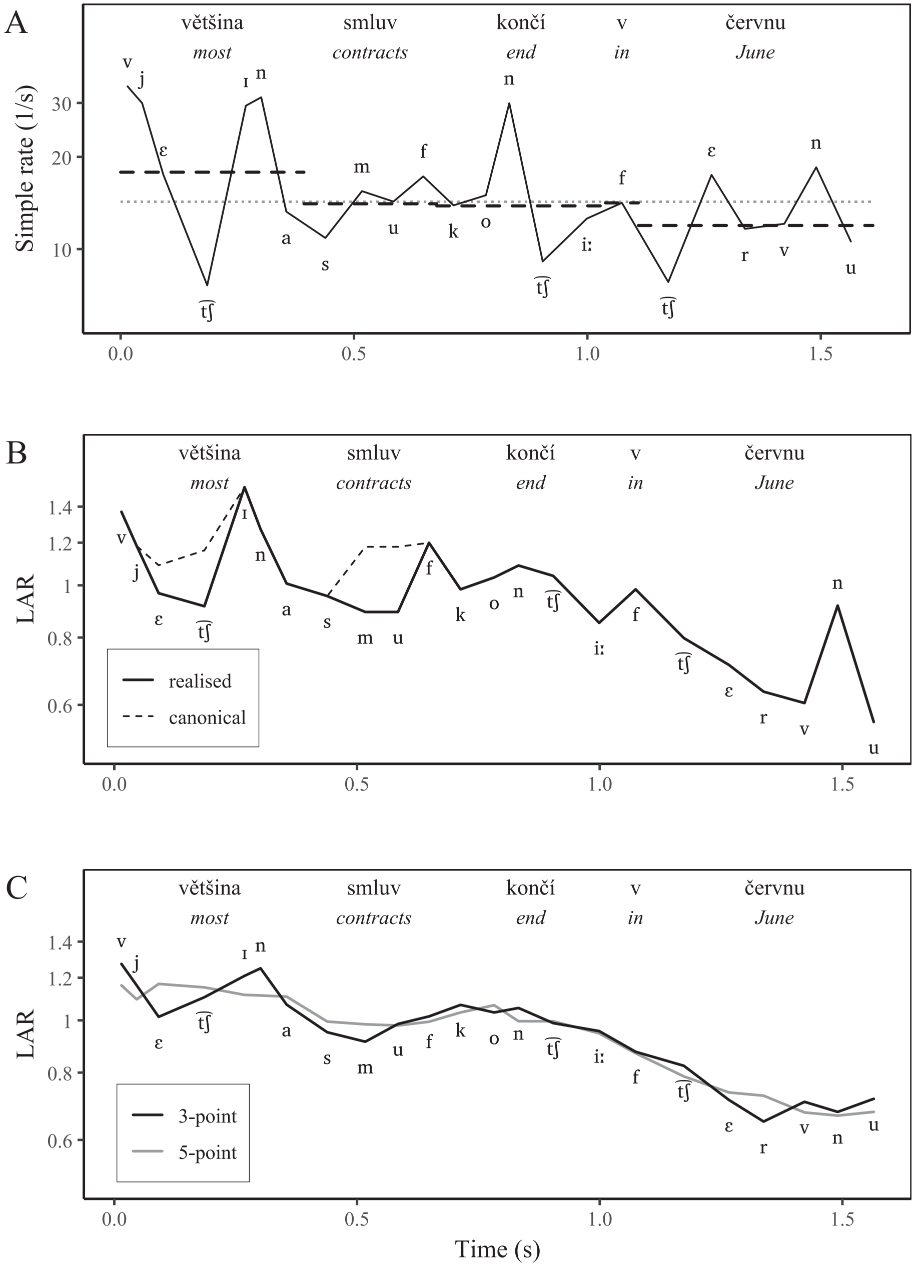

Figure 1 The simple rate and local articulation rate (LAR) contours of an example phrase. Panel A: mean articulation rate of the phrase (dotted line), simple articulation rate in words (dashed line) and by phones (solid line). Panel B: realised (solid line) and canonical (dashed line) local articulation 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 [t͡ʃiː] 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.

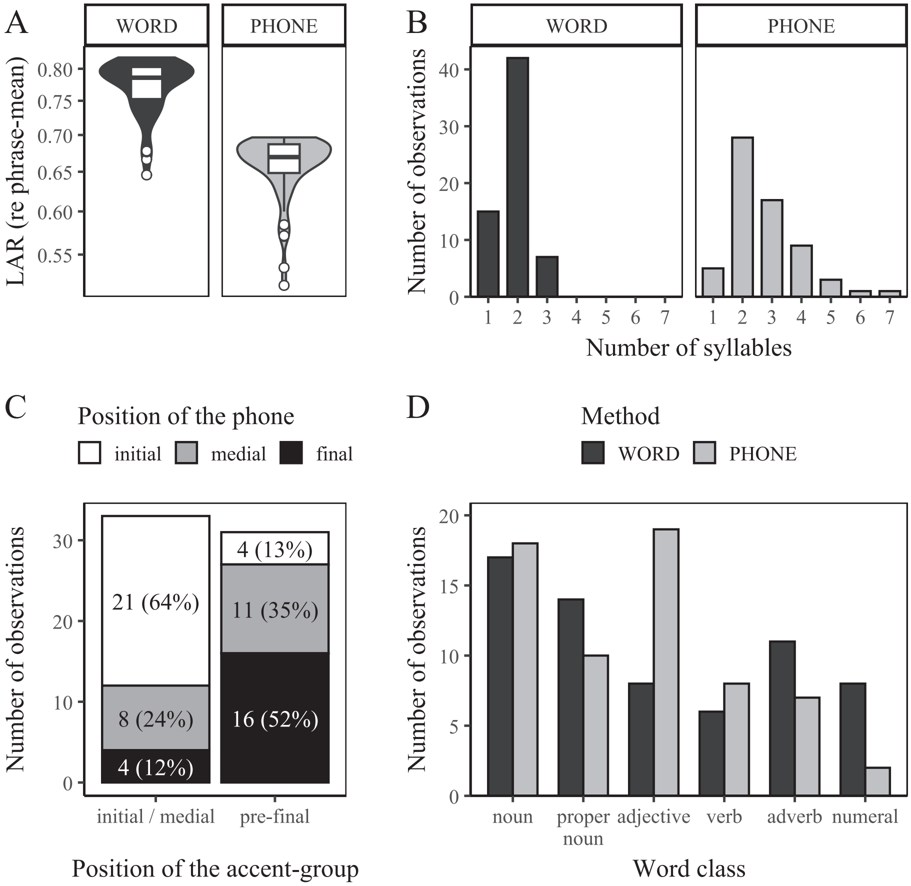

Figure 2 Summary statistics of the selected decelerated content words identified by the WORD and PHONE methods (see text for explanation). Panel A: local articulation rate (LAR) values of the target word or slowest phone. Panel B: target word length (in syllables). Panel C: position of the slowest phone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

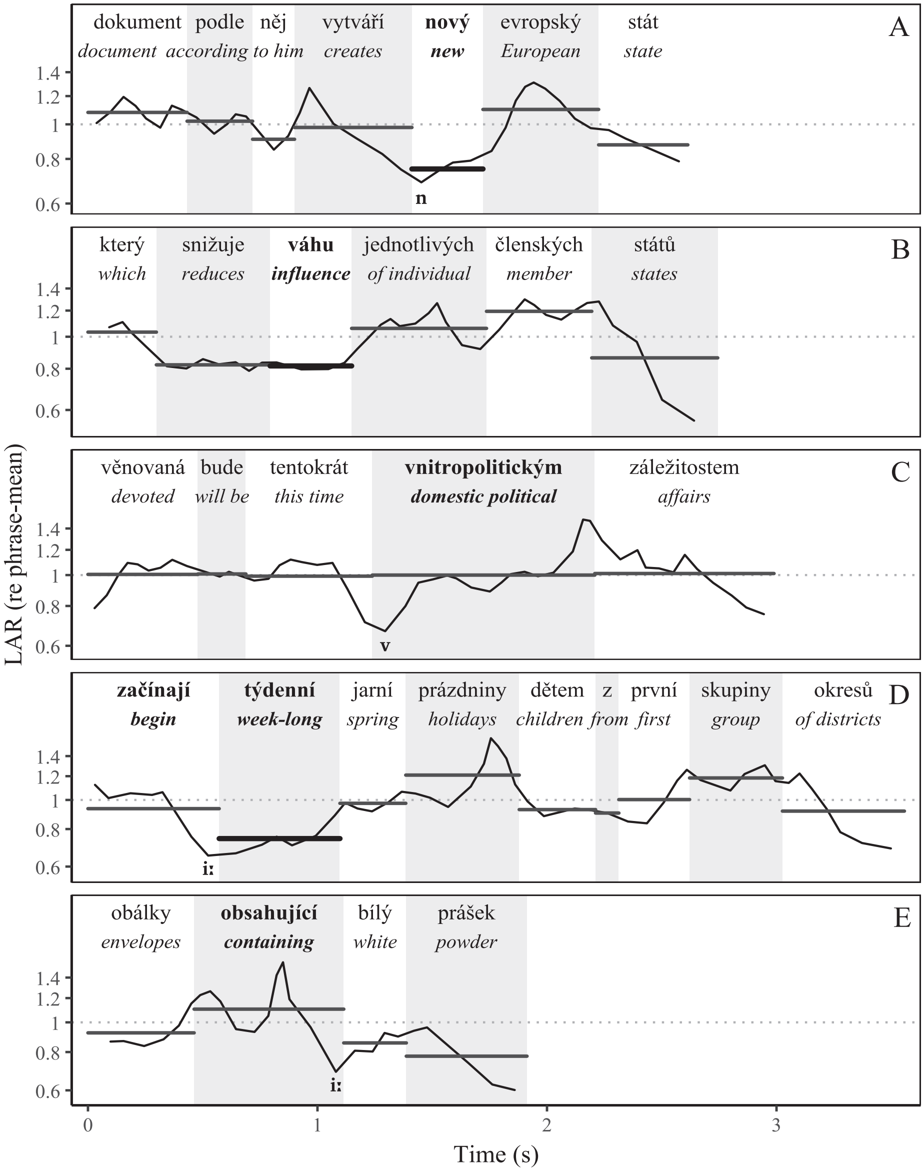

Figure 3 The local articulation rate (LAR) contours (smoothed with a 3-point moving average) of selected phrases containing locally decelerated words. The horizontal lines represent LAR means in words; the decelerated target words are highlighted in bold. For wider contexts and interpretations (especially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

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%)
			contrastive topic	3 (5%)
	rheme	46 (72%)	part of rheme	37 (58%)
			rheme (contrast)	1 (2%)
			rheme proper	8 (12%)
PHONE	theme	23 (36%)	theme proper	---
			part of theme	21 (33%)
			contrastive topic	2 (3%)
	rheme	41 (64%)	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-

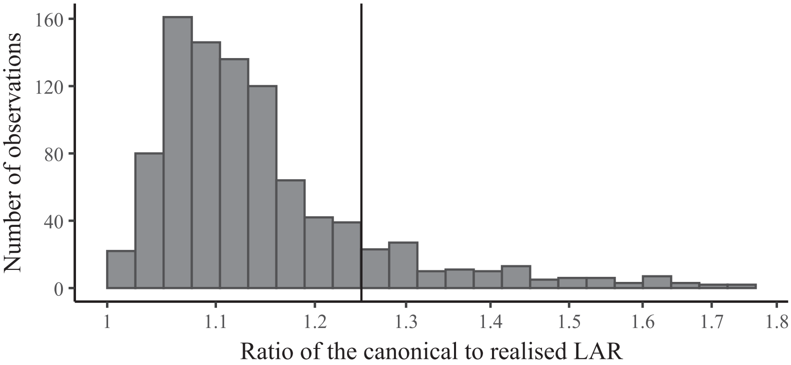

Figure 4 The histogram of the differences between the two local articulation rate (LAR) measures, expressed as the ratio of the canonical to the realised mean LAR in words. For 16% of words, the ratio was higher than 1.25 (indicated by the vertical line).

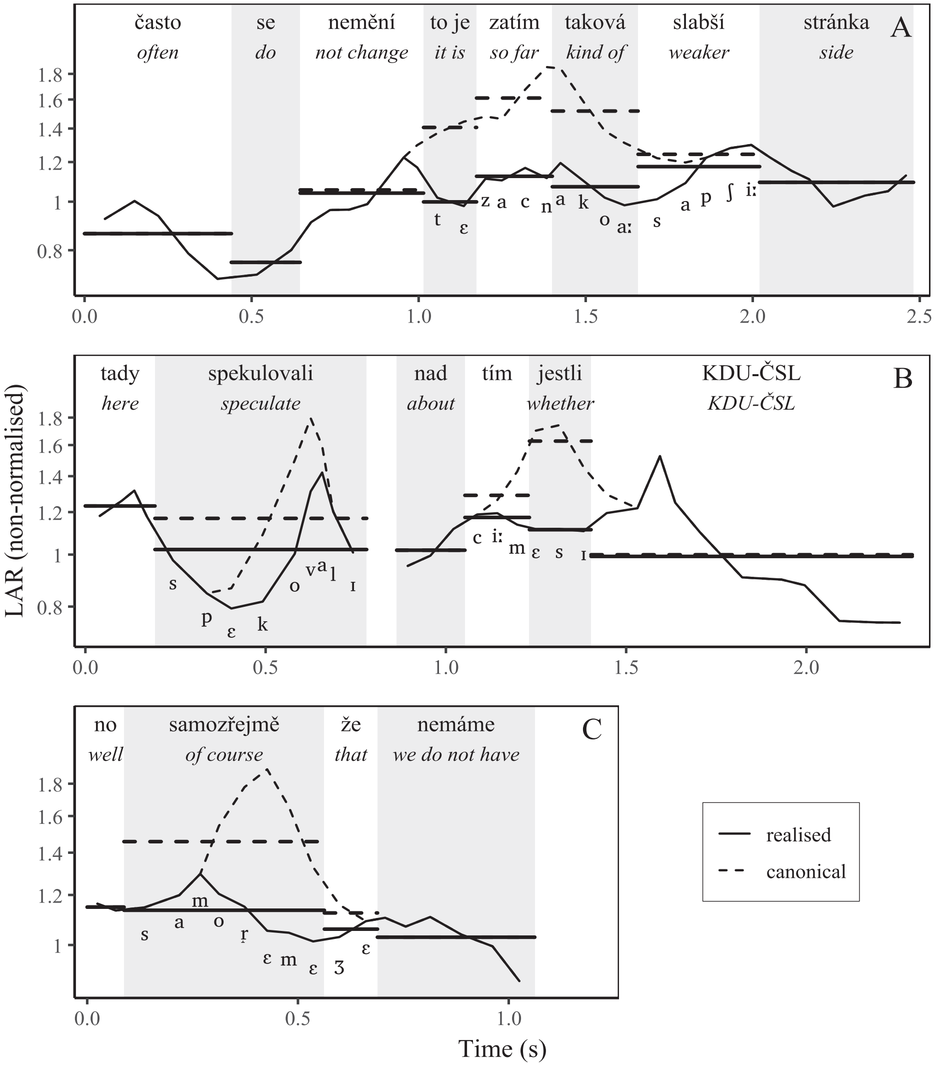

Figure 5 The realised (solid line) and canonical (dashed line) local articulation rate (LAR) contours of selected phone reductions, smoothed with a 3-point moving average. The horizontal lines represent realised and canonical LAR means in words.

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