Examining Temporal Structure of Speech with a Local Articulation Rate Metric

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

Michaela Svatošová
Institute of Phonetics, Charles University, Prague
michaela.svatosova@atarien.com

Abstract
The primary goal of our study is to propose a method of calculating and visualising local articulation rate for research in temporal structure of speech. The method builds on proportional durations of vowels and consonants in Czech, which normalizes for inherent durations of phones. We first demonstrate the importance of temporal structure on several conspicuous features: phrase-final deceleration, prominence marking, parentheticals, and information structure constituents. We then describe our method stepwise so that it could be tested by interested parties. We illustrate such testing on a sample of news bulletin sentences produced by 26 speakers. The results confirm that our procedure can meaningfully reflect various temporal features, including the 'information status' of words in contextually grounded utterances.

Keywords: final deceleration, local articulation rate, information structure, information status, phone duration, temporal contour.

1. Introduction

1.1. State of the art in the research area

The importance of temporal structure of speech is generally acknowledged at present, even if not fully appreciated. It is intuitively accepted, for instance, that excited or eagerly involved people talk faster than those who are bored or hesitant. There are studies that found differences in speaking rates among speech styles or communicative genres (although their claims are sometimes contradicting). We also know that emphasis makes some words slower while contextually given words are pronounced faster. In other words, we suppose that differences in local articulation rate can be used for coding all three components of utterance meanings: referential, conative, and affective (Bühler, 1934; cf. also Volín, 2019: 126). A pressing task in current research is to find out how this is done.

One of the best-studied temporal events in speech is the phrase-final deceleration (phrase-final lengthening). Its significance was underscored in a corpus study of German (Peters, Kohler, & Wesener, 2005). It indicated that while pauses were rarely present as the only boundary cue, deceleration and final nuclear melodic patterns occurred in this role much more frequently. The impact of final deceleration on comprehension or intelligibility is often mentioned even in studies where it is not quantified (Martin, 1968; Price, Ostendorf, & Shattuck-Hufnagel, 1991; Ferreira, Anes, & Horine, 1996; Carlson, Clifton, & Frazier, 2001; Hirotani, Frazier, & Rayner, 2006; Breen, Watson, & Gibson, 2011).

On the other hand, the studies that focused on quantitative characteristics yielded disparate results even within a single language (e.g., Scott, 1982 vs. Aasland & Baum, 2003 for English, or Zhang, 2012 vs. Yang, Shen, & Yang, 2014 for Mandarin). The results are not easily comparable due to methodological differences – for example, maximal duration of a pause was 80 ms in Zhang (2012) and over 500 ms in Scott (1982). Therefore, as noted by Yang, Shen, & Yang (2014), it is desirable to work on developing compatible approaches.

We found a similar situation in another area of interest in temporal structure: the issue of parenthetical constituents or insertions in sentences. Speakers use them to host secondary information supplemented to the main topic. Analysing conversational speech, Local (1992) reported a higher speech rate during the insertion itself, while a switch to a slower tempo signalled the return to the main idea. However, Lelandais and Ferré (2014) tested his claim with their own material and found that some types of insertions were faster, while other types were slower than the surrounding context. Nevertheless, some change in tempo was common to all of them.

A related problem to the above is that of information structure. Language users typically produce utterances which, on the one hand, evoke certain concepts known to them and the recipients of their message, and on the other hand, contain some new information on those concepts. Even though this is not the only possible configuration, it seems to be the prevailing one. Felicitousness of communication (the term after Austin, 1962 and Searle 1969, 1979; *cf.* also Grice, 1991, p. 19) can thus depend on the correct signalling of what the recipient is expected to know and what is supposed to update his/her knowledge.

In research, the identification of the known (related to concepts of *theme*, *given* or *inferable*) and the new (related to *focus* or *rheme*) becomes a task of information structure analysis (also discussed as Functional Sentence Perspective – FSP, e.g., Firbas, 1992).

Prosodic marking of the new or the given has been reflected for decades. The new usually bears special prominence, while the given or inferable is prosodically backgrounded: slower speech rate was found in phrases that consisted of new or important information, whereas higher speech rate was used

for already mentioned facts and in self-corrections (Uhmann, 1992). Multiple studies have explored the relation of temporal cues and information structure (e.g., Cooper, Eady & Mueller, 1985 for English, Baumann et al., 2007 and Kügler, 2008 for German, and Heldner & Strangert, 2001 for Swedish). All of them found a consistent pattern – words have longer durations when they are in narrow or contrastive focus as opposed to broad focus. Moreover, the position of a word in an utterance seems to have an effect on the amount of duration increase since focal lengthening interacts with deceleration in phrase-final position. (Nevertheless, various researchers measuring this parameter found disparate results regarding the direction of the difference). The variability might also be affected by speaker preferences (Heldner & Strangert, 2001, p. 355).

The typical design of the above-mentioned studies relied on read sentences elicited by questions that ensure the intended placement of focus. Moreover, the prominence or its absence was often just stated but not phonetically specified. Even the researchers who care to describe prominence marking usually refer to speech melody (cf. Kohler, 2006; Baumann & Riester, 2013; Büring, 2019). One of the goals of our current article is to encourage the research community to use our method with temporal data in order to expand scientific knowledge of prominence marking and other prosodic features just mentioned. (Or, alternatively, to discuss its shortcomings and suggest improvements).

1.2. The current research objectives

When studied, temporal features of utterances are usually expressed through global parameters, e.g., mean articulation or speech rate over larger stretches of speech. Although such approaches often lead to useful and interesting findings, we believe that a more fine-grained investigation of temporal forms is also needed. Mean tempi over prosodic phrases (intonation phrases, tone-units) can hardly reveal their inner temporal structure. On the other hand, the interpretation of deceleration as individual phone lengthening, although mathematically correct, obscures its prosodic foundation. Speakers' language competences fit better into the existing prosodic framework if we conceptualize them as slowing down (decelerating) or speeding up (accelerating) on the words/morphemes that are used, rather than as stretching or shrinking individual phoneme representations that carry no meaning. A methodological obstacle to avoid these two problems rests in the difficulties with operationalizing local articulation rate (LAR). Therefore, our current objectives could be specified as follows.

First, we would like to propose a procedure that is capable of measuring the local articulation rate (LAR) in utterances without the bias produced by disparate inherent durations of phones. The procedure can be also used to produce LAR

contours that visualise the temporal flow similarly to, for instance, fundamental frequency tracks produced by F0 extractors. Moreover, the method should also allow for estimating LAR in individual utterance constituents, i.e., to display on which words the speaker either accelerated or decelerated.

Second, we wish to demonstrate how the proposed procedure works in temporal structure examination. A sample of continuous spoken texts with a certain communicative intent is used. Unlike laboratory sentences, our material presents a more ecologically valid use of language, even though it is at the expense of control over the speakers' behaviour. This makes the test of our procedure more stringent.

2. Method

2.1. Material

Three types of speech material were used in this study: (1) news reading (NWS) and (2) story telling (STR) for the development of the local articulation rate procedure, and (3) duplicated news bulletins (DNB) for the illustrative temporal structure analyses.

To develop the LAR metric, we used speech production of 16 radio news readers and 16 audio books narrators (8 female & 8 male speakers in each genre). All the speakers were experienced professionals, renowned in their career fields. The news bulletins (NWS) were broadcast by Czech Radio I and II (official national broadcaster), and the stories (STR) were produced in professional studios to be sold as audio books. This NWS and STR material comprised 25,353 words, which gave 124,236 individual phone tokens representing 39 Czech phonemes.

The DNB material was intended for analyses that were to verify the usefulness of the LAR metric. It comprised 26 renderings of a news bulletin text originally broadcast by a national radio station. We hired 26 speakers (14 female + 12 male) to record it for us. All of them were current or former university students majoring in philological programmes, without speech disorders or hearing problems, with Czech as their mother tongue, and they ranged between 19 and 33 years of age.

The speakers were given texts of individual news paragraphs on separate sheets of paper and were asked to get familiar with the contents. They were encouraged to practice the performance, but they were also reassured that any mistakes would be edited out and their performance would be strictly anonymous. All recordings were made in the sound-treated studio of the Institute of Phonetics in Prague. (No ethics approval was necessary.) The studio was equipped with a condenser microphone AKG C4500 B-BC plugged directly into an external soundcard Steinberg UR44. The signal was saved with 32-kHz sampling rate and 16-bit resolution in an uncompressed format as wav files.

The actual text originally comprised six paragraphs plus some introductory and concluding phrases. For the purpose of the current analysis, one sentence from paragraphs 2, 3, and 4 was extracted. These sentences were clearly contextually anchored as they were always taken from the second half of the respective

paragraph. (At least two other sentences preceded them.) The actual content of the sentences and the contextual grounding is presented below in Section 3 – Results. All recordings were processed identically. Forced alignment for words and phones was performed with Prague Labeller (Pollák, Volín & Skarnitzl, 2007), and manual corrections through auditory and visual inspection were carried out in Praat (Boersma & Weenink, 2022).

2.2. The LAR metric

The conceptual foundation of local articulation rate (LAR) rests in the idea that duration of a phone can be inverted into a rate by

$$AR_{pho} = \frac{1}{dur_{pho}} \tag{1}$$

where dur_{pho} is duration of a segment and AR_{pho} is its articulation rate. However, the inherent durations of various phones would introduce bias, by which inherently short phones (e.g., intervocalic [r], [l], [d]) would mistakenly suggest local acceleration, while long phones (e.g., [a:], [o͡u], [t͡ʃ]) would imply deceleration. Thus, words with inherently long phones would mislead an observer to believe that they are slow, and vice versa.

To avoid this bias, we computed durational ratios (DRs) of phones relative to vowel [e]. (Mean duration of [e] in our corpus was 54 milliseconds.) This vowel was elected due to its highest frequency of occurrence and, also, due to its very consistent durational behaviour: its coefficient of variation was between 20 and 30% for the individual speakers in the corpus. (According to Bartoň et al., 2009, [e] is also the most frequent phone in Czech.) The formula is as follows:

$$DR_{pho} = \frac{\bar{d}_{[e]}}{\bar{d}_{pho}} \tag{2}$$

where DR_{pho} is the durational ratio of a phone, $\bar{d}_{pho}$ is mean duration of a phone and $\bar{d}_{[e]}$ is the mean duration of [e] in the corpus. (The reason why $\bar{d}_{[e]}$ is in numerator and $\bar{d}_{pho}$ in denominator is in more straightforward calculation of the next steps in the procedure.)

Durational ratios are based on mean phone durations in our large data set after exclusion of phones in phrase-final syllables and syllable rimes of pre-final positions (the domain of phrase-final deceleration). Furthermore, we excluded post-pausal plosives and affricates since for these it is impossible to establish where their articulations actually starts. Finally, we excluded a tiny fraction of the data (0.14 %) constituted by under-represented marginal elements of the Czech phone inventory (representations of imported phonemes /o:/, /a͡u/, /e͡u/, /d͡ʒ/, facultative phones [m̩], [ŋ], and rare allophones [ɣ], [d͡z]). Thus, the original pool of 124,236 tokens was reduced to 102,026 tokens. (Analogically, 12,568 tokens of [e] were reduced to 10,040 tokens after the exclusion of final and prefinal cases.)

To avoid pseudo-replication in our data, we calculated the DRs for each speaker first, and only after that the grand mean across speakers was produced. Figure 1 and Table 1 display the resulting ratios. It can be observed that most phones had mean duration longer than that of [e], which is signalled by durational ratios lower than one. Figure 1 also shows that short vowels display very small variation across speakers. (Short vowels in Czech spoken texts are about four times more frequent than the long ones.)

Table 1: Durational ratios obtained from phone means relative to mean duration of [e].

phone	durational ratio	phone	durational ratio	phone	durational ratio
a	0.89	ŋ	0.85	g	0.82
e	1.00	r	1.10	f	0.71
ɪ	1.04	r̥	0.79	v	1.02
o	1.00	l	1.19	s	0.54
u	0.99	l̥	0.82	z	0.71
a:	0.45	j	1.13	ʃ	0.53
e:	0.57	p	0.62	ʒ	0.71
i:	0.76	b	0.78	x	0.62
o͡u	0.50	t	0.73	ɦ	0.89
u:	0.69	d	0.99	t͡s	0.47
m	0.82	c	0.63	t͡ʃ	0.45
n	1.01	ɟ	0.79	r̥̊	0.70
ɲ	0.94	k	0.64	r̥	0.80

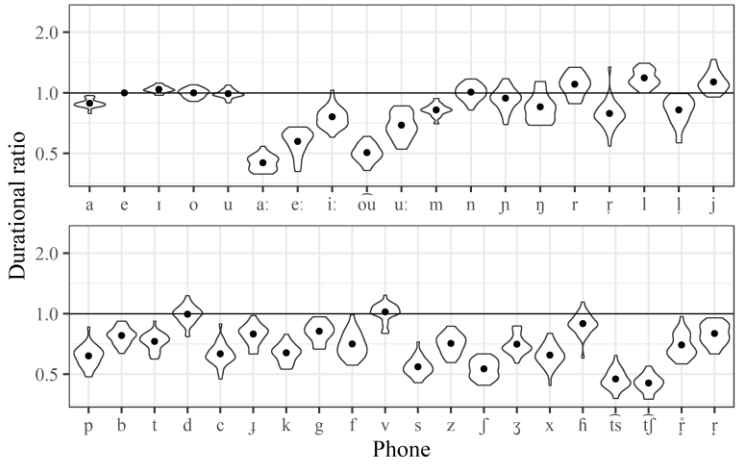

Figure 1: Durational ratios obtained from mean phone durations. The violin plots reflect distribution of individual speakers' ratios; the dots are grand means (averaged ratios).

The real measured durations in actual utterances were then multiplied by mean durational ratios (DRs) obtained from our large set of data by the following formula:

$$dur_{norm} = dur_{pho} * DR_{pho} \quad (3)$$

where dur_{norm} is the normalised duration of a phone, dur_{pho} is the measured duration in an utterance and DR_{pho} is the durational ratio of the given phone (from Table 1).

The normalised durations were inverted into local articulation rates by Formula 1. This procedure still produced quite uneven LAR contours, which was counter-intuitive in that listeners are not expected to re-evaluate perceived local articulation rate from phone to phone. Therefore, the third step that we propose is smoothing the contour by a moving average. Heuristically, we opted for a three-point moving average.

Table 2: Example of step by step calculations described in Section 2.2. The first five phones of the illustrative sentence from Figure 2 below were taken.

phone	measured duration (seconds)	durational ratio (DR)	normalised duration (seconds)	local articulation rate (LAR)	smoothed LAR
k	0.081	0.64	0.052	19.3	16.9
a	0.078	0.89	0.069	14.6	17.3
r	0.050	1.10	0.055	18.1	16.3
m	0.075	0.82	0.062	16.2	16.2
i:	0.093	0.76	0.071	14.1	16.5
etc.	…	…	…	…	…

Finally, as an alternative to the 'moving-average' contour, we also calculated mean normalized LAR for individual words since words are the constructional elements that are typically considered in language analyses. Figure 2 shows the raw rates and the ratio-normalized rates in panel A, whereas the smoothed contour and word averages are in panel B. (The utterance in the figure is just an illustrative example from an existing narrative, i.e., the STR corpus.) The software used in our case was *R Studio* (R Core Team, 2022) with packages *tidyverse* (Wickham, 2019) and *rPraat* (Bořil & Skarnitzl, 2016).

In summary, our procedure comprises the following steps:

a) establishing mean durations of phones from a large representative corpus
b) calculating durational ratios (DRs) of individual phones relative to the most frequent vowel for each speaker separately (Formula 2)
c) averaging the speakers' DRs to produce the mean DRs for the corpus
d) computing normalized durations of phones in utterances (Formula 3)
e) inverting the normalised durations of phones into local articulation rates (LARs, Formula 1)

f) smoothing the sequences of durations in utterances with 3-point moving average
g) visualizing the LAR contour
h) optional but recommended: calculating mean LAR for individual words

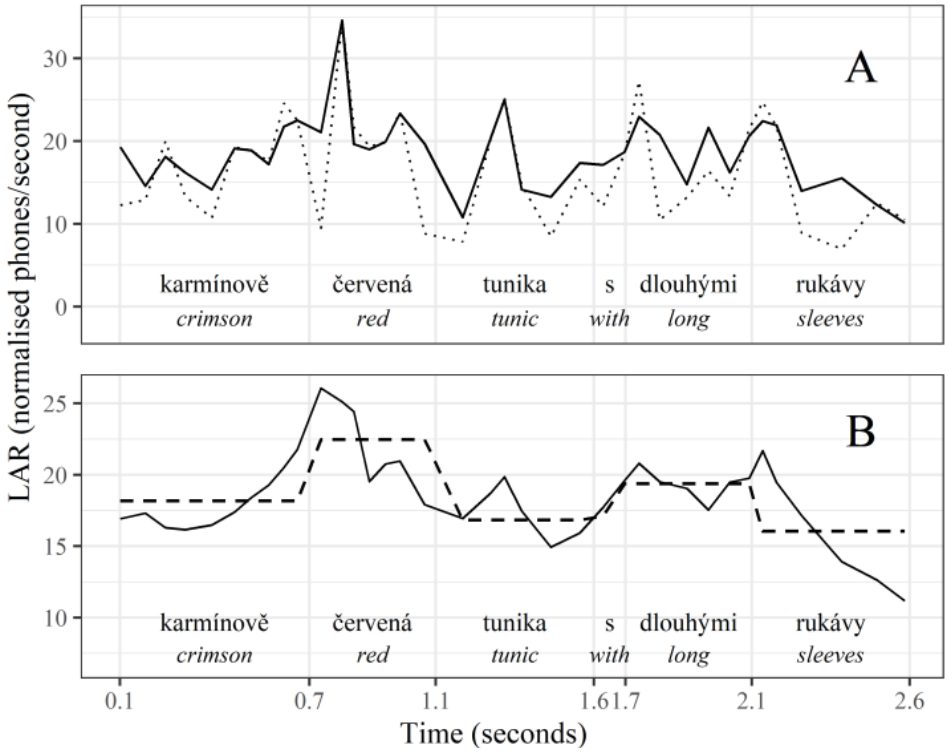

Figure 2: Illustration for the development of the LAR metric: Panel A: (1) raw values (dotted line), (2) values normalised by the DRs (solid line); Panel B: (3) smoothing by three-point moving average (solid line), (4) means for words (dashed line).

In the following section, the results of our probe into the DNB material is presented. The DNB material (see also Section 2.1) represents the same speech style as the corpus that served for the foundation of our durational modelling, but it was not used in the calculations themselves. We intentionally test our procedure on material that was not part of the procedure development.

3. Results

The outcomes of the LAR measurements on the three selected sentences (produced in texts spoken by 26 speakers specified in Section 2.1) are displayed in Figure 3, panels A, B, and C. Each panel contains lines that represent three

conditions of LAR extraction. First, temporal contours could be averaged phone-wise and word-wise (dotted line and solid line, respectively). Additionally, we inspected performance of individual speakers with respect to their reading skills: eight speakers were selected that seemed to be most 'present' when reading the text. The 'presence' is an impressionistic evaluation of the speech performance made by the authors of this study. (It is opposite to 'mechanistic' reading which gives impression of detached, automatized language use.) The 'present' speakers seem to care about the contents of the story they are reading out. Theirs is the third temporal contour – the solid one in Figure 3.

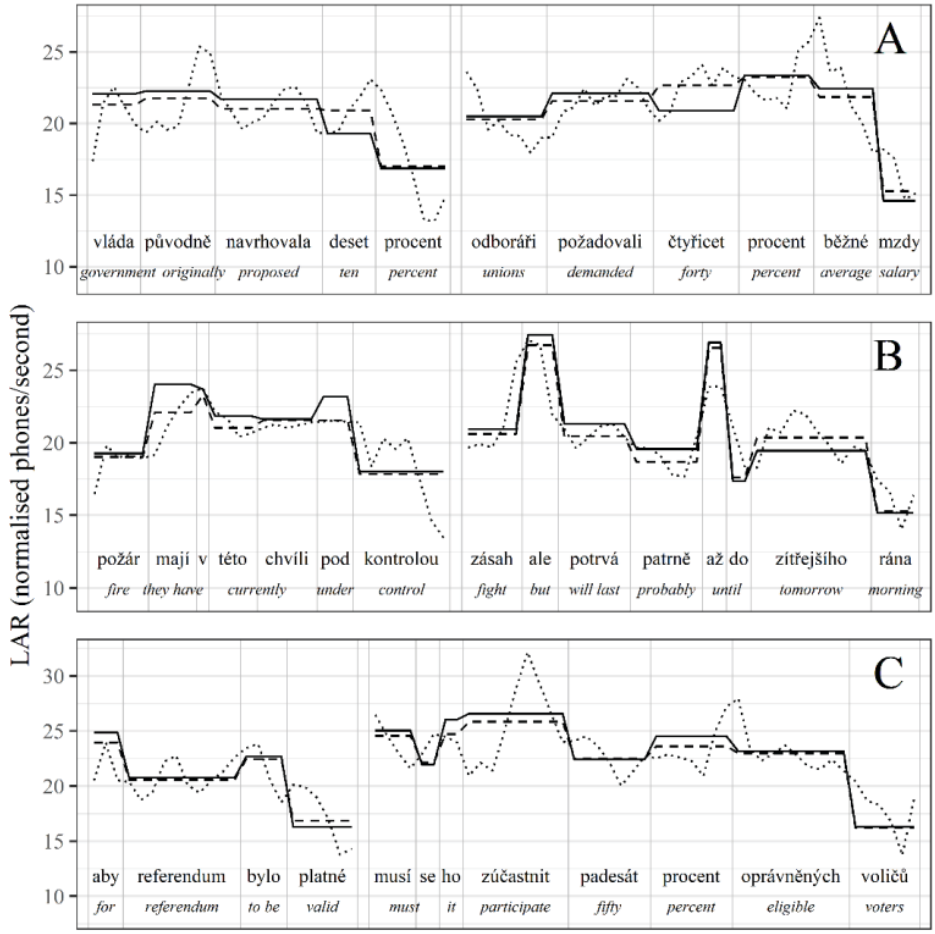

Figure 3: Three example utterances, each with three outcomes of the presented procedure: (1) the dotted line = mean smoothed (three-point average) normalised LAR contour; (2) the dashed line = mean normalised LAR contour (means for words); (3) the solid line = the mean normalised LAR contour (means for words) for eight selected speakers (see text).

It can be observed that all three utterances clearly split into two prosodic phrases with the first phrase shorter than the second one (in line with Behaghel's rule, Behaghel, 1909). As expected, the phone based LAR contour (dotted in Fig. 3) is the most ragged one, but it allows for more detailed segmental analysis. However, since the word is considered the basic building block of utterances, the word-based LAR contour (dashed in Fig. 3) might be of greater interest to linguists even if it conceals certain detail. For instance, all three types of visualisation clearly display quite distinct phrase-final deceleration: the last words are always the slowest ones. Nevertheless, the dotted phone-based line shows that this is happening gradually and the very last phone in a phrase might occasionally behave inconsistently (→ a matter of future research).

Another observation valid for all three utterances is that the selected speakers were slightly more contrastive: they reached higher and usually also lower LARs in an utterance (solid line in Fig. 3). Utterance specific observations are discussed below.

3.1. Utterance A

The wording of the utterance in Czech is *Vláda původně navrhovala deset procent, odboráři požadovali čtyřicet procent běžné mzdy.* (In English: *The government originally proposed ten percent, trade unions demanded forty percent of the average salary.* The English translation word by word is in Fig. 3 under the contour A.) The preceding co-text introduced the fact that the Czech government and trade unions had met to negotiate salaries of state employees. The highest LAR is on the second repetition of the word *procent* (*percent*). The first usage of the word at the end of the first phrase is very slow, chiefly because it is in the domain of phrase-final deceleration. What needs to be highlighted is the fact that relative to the sample average, the selected speakers slow down on words *deset* (*ten*) and *čtyřicet* (*forty*). These numeral expressions are the informational core of the utterance and they are also uninferable: the listeners cannot guess how much money the government offered and how much the unions demanded. Their information status is, therefore, high. (Uninferability is termed 'high communicative dynamism' in Firbas, 1992.) The word *vláda* (*government*) is quite fast – it is already known from the immediately preceding sentence. So is the word *odboráři* (*unions*) at the beginning of the second phrase, but its status is slightly different because here it is put in contrast to the government. Contrastive focusing is one of the prominent topics in information structure research.

3.2. Utterance B

The wording of the utterance in Czech is *Požár mají v této chvíli pod kontrolou, zásah ale potrvá patrně až do zítřejšího rána*. (In English: *The fire is currently under control, but the firefighting will probably go on until tomorrow morning.*)

Although this is already the fourth sentence in a paragraph dedicated to quite a damaging fire and its extensive fighting by several teams, the word *požár* (*fire*) was only mentioned at the very beginning, in the first sentence. The return to *požár* in the position of the utterance topic in this fourth sentence apparently required slower articulation (lower LAR). The auxiliary verb *mají* (*have*) with the *inflection* signalling *they* is noticeably accelerated, given its informationally week status. The expression *v této chvíli* (*currently*) is relatively fast because of its predictable deictic function. In the second phrase, there are two LAR peaks on grammatical words: *ale* (*but*) and *až* (*until*). Although Czech has always been referred to as an unreducing language, here we can see that grammatical operators are pronounced markedly fast. The rhematic expression *zítřejšího rána* (*tomorrow morning*) is, on the other hand, slow. In addition, *rána* (*morning*) hosts phrase-final deceleration, but this is often limited to the last syllables, so the word before can be considered unaffected, i.e., hosting only the informational deceleration. (Obviously, these observations are only illustrative, but they can easily serve to stipulate hypotheses for future research.)

3.3. Utterance C

The wording in Czech is *Aby referendum bylo platné, musí se ho zúčastnit padesát procent oprávněných voličů.* (In English: *For the referendum to be valid, fifty percent of eligible voters must participate*.) The behaviour of the selected speakers in this utterance does not differ much from that of the whole sample. Yet, there are still some relevant moments to observe. The first phrase has two fast and two slow words. The fast ones are the conjunction *aby* (*so that*) and the copula *bylo* (*was*) – clearly just grammatical operators. The slow words are autosemantic: the noun *referendum* (*referendum*) and complementing adjective *platné* (*valid*). The latter one also acts as the phrase-final deceleration domain.

The fastest word in the second phrase is the verb *zúčastnit* (*participate*). It is highly inferable, since referendum implies participation. An informationally relevant difference can be seen between *padesát* (*fifty*) and *procent* (*percent*). The word *percent* can be inferred, hence it is faster, while *fifty* cannot, so it is slower. The subset of selected speakers (solid line in Fig. 3) made the difference between inferable and non-inferable lexical items larger.

4. Discussion

The importance of the prosodic structure of speech is undisputed. Its impact is supported by evidence from many areas of phonetic research and, for instance, even from neurolinguistics owing to multiple experiments studying the Closure Positive Shift, which is a specific event related potential (ERP) of brain activity connected to prosodic boundary perception (e.g., Holzgrefe-Lang et al., 2016).

Although our proposed method has been tested on only a limited sample, it seems to serve its purpose. The resulting contours of local articulation rate are compatible with the current general knowledge of phrase-final deceleration, prominence marking, and information structure tenets.

However, discussion by the research community is essential. There is definitely space for improvement but, ideally, the objections against the proposed method will be based on its use under various conditions with diverse speech material.

Two serious limitations that we are aware of can be mentioned already. First, the durational ratios, although established from more than 100,000 phones, represent very similar articulation rates (ranging between 6 and 7 syll/sec.). The ratios quite probably blur the situation in slower or faster rates, since it is known that the temporal elasticity of vowels is markedly higher than that of consonants. The extent of distortion needs to be examined and if found significant, rectification must be devised.

Second, our 'clear speech' corpus did not have to deal with elisions and parallel articulations. However, conversational speech harbours those quite frequently and the treatment of such 'zero durations' needs to be handled in an informed manner, which is currently unavailable to us.

A tempting task would be to ground the normalization in syllabic rather than phone durations. However, Czech with its relatively free phonotactics of consonant clusters allows for thousands of different syllables (Šturm & Bičan, 2021). Nevertheless, the phone-based LAR could be used to detect durational anomalies connected with syllable formation and, in extension, in research of the role of syllable in articulatory planning.

Even though the observation made throughout our testing (described in Section 3) seem to be compatible with the current views on the information structure of utterances, they have to be verified in dedicated, carefully designed experiments before they are generalized. Only too often it has been discovered in linguistics that what sounds logical is actually empirically untenable.

The visualization of LAR can be useful in any area where the temporal forms are of interest and, at the same time, many things need to be investigated further. At the moment, we are able to produce temporal contours that already seem to be functional in the research of language use. However, their further development and rules of their interpretations will depend on future work. We hope that our proposal can serve to further advancement of human understanding of speech communication.

Acknowledgements

This research was supported by the Grant Agency of the Czech Republic, project 21-14758S.

References

Aasland, Wendi A. and Shari R. Baum. 2003. Temporal parameters as cues to phrasal boundaries: A comparison of processing by left- and right-hemisphere brain-damaged individuals. *Brain and Language*, 87(3), 385–399. https://doi.org/10.1016/S0093-934X(03)00138-X

Austin, John L. 1962. *How to Do Things with Words*. Oxford University Press.

Bartoň, Tomáš; Cvrček, Václav; Čermák, František; Jelínek, Tomáš and Vladimír Petkevič. 2009. *Statistiky češtiny*. Nakladatelství Lidové noviny.

Baumann, Stefan and Riester, Arndt. 2013. Coreference, lexical givenness and prosody in German. *Lingua*, 136, 16-37. https://doi.org/10.1016/j.lingua.2013.07.012

Baumann, Stefan; Becker, Johannes; Grice, Martine and Doris Mücke. 2007. Tonal and articulatory marking of focus in German. In *Proceedings of the XVIth ICPhS*, 1029-1032.

Behaghel, Otto. 1909. Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. *Indogermanische Forschungen*, *25*, 110-142. https://doi.org/10.1515/9783110242652.110

Boersma, Paul and David Weenink. 2022. *Praat: Doing Phonetics by Computer* (6.2.07). https://www.praat.org/

Bořil, Tomáš and Radek Skarnitzl. 2016. Tools rPraat and mPraat. In P. Sojka, A. Horák, I. Kopeček and K. Pala (eds.), *Text, Speech, and Dialogue, Lecture Notes in Computer Science*, 367-374. Springer International Publishing. https://doi.org/10.1007/978-3-319-45510-5_42

Breen, Mara; Watson, Duane G. and Edward Gibson. 2011. Intonational phrasing is constrained by meaning, not balance. *Language and Cognitive Processes*, 26(10), 1532-1562. https://doi.org/10.1080/01690965.2010.508878

Bühler, Karl. 1934. *Sprachtheorie*. Fischer. (currently available in the English version: *Theory of Language*, 2011, John Benjamins Publishing Company)

Büring, Daniel. 2019. Focus, questions and givenness. In K. von Heusinger, E. Onea and M. Zimmermann (eds.), *Questions in Discourse*, 6-44. Brill.

Carlson, Katy; Clifton, Charles and Lyn Frazier. 2001. Prosodic boundaries in adjunct attachment. *Journal of Memory and Language*, 45(1), 58-81. https://doi.org/10.1006/jmla.2000.2762

Cooper, William E.; Eady, Stephen J. and Pamela R. Mueller. 1985. Acoustical aspects of contrastive stress in question–answer contexts. *The Journal of the Acoustical Society of America*, 77(6), 2142-2156. https://doi.org/10.1121/1.392372

Ferreira, Fernanda; Anes, Michael D. and Matthew D. Horine. 1996. Exploring the use of prosody during language comprehension using the auditory moving window technique. *Journal of Psycholinguistic Research*, 25(2), 273-290. https://doi.org/10.1007/BF01708574

Firbas, Jan. 1992. *Functional Sentence Perspective in Written and Spoken Communication*. Cambridge University Press.

Grice, Paul. 1991. *Studies in the Way of Words*. Harvard University Press.

Heldner, Mattias and Eva Strangert. 2001. Temporal effects of focus in Swedish. *Journal of Phonetics*, 29(3), 329-361. https://doi.org/10.1006/jpho.2001.0143

Hirotani, Masako; Frazier, Lyn and Keith Rayner. 2006. Punctuation and intonation effects on clause and sentence wrap-up: Evidence from eye movements. *Journal of Memory and Language*, 54(3), 425-443. https://doi.org/10.1016/j.jml.2005.12.001

Holzgrefe-Lang, Julia; Wellmann, Caroline; Petrone, Caterina; Räling, Romy; Truckenbrodt, Hubert; Höhle, Barbara and Isabell Wartenburger. 2016. How pitch change and final lengthening cue boundary perception in German: converging evidence from ERPs and prosodic judgements. *Language, Cognition and Neuroscience*, 31(7), 904-920. https://doi.org/10.1080/23273798.2016.1157195

Kohler, Klaus J. 2006. What is emphasis and how is it coded? In *Proceedings of the 3rd International Conference on Speech Prosody*, 748-751.

Kügler, Frank. 2008. The role of duration as a phonetic correlate of focus. In *Proceedings of the 4th Conference on Speech Prosody*, 591-594.

Lelandais, Manon and Gaëlle Ferré. 2014. Multimodal analysis of parentheticals in conversational speech. *Multimodal Communication*, 3(2). https://doi.org/10.1515/mc-2014-0008

Local, John. 1992. Continuing and restarting. In P. Auer and A. Di Luzio (eds.), *Pragmatics & Beyond New Series*, 273-296. John Benjamins Publishing Company. https://doi.org/10.1075/pbns.22.18loc

Martin, James G. 1968. Temporal word spacing and the perception of ordinary, anomalous, and scrambled strings. *Journal of Verbal Learning and Verbal Behavior*, 7(1), 154-157. https://doi.org/10.1016/S0022-5371(68)80181-1

Peters, Benno; Kohler, Klaus and Thomas Wesener. 2005. Phonetische Merkmale prosodischer Phrasierung in deutscher Spontansprache. Prosodic Structures in German Spontaneous Speech (AIPUK 35a).

Pollák, Petr; Volín, Jan and Radek Skarnitzl. 2007. HMM-Based Phonetic Segmentation in Praat Environment. In *The XII International Conference Speech and Computer – SPECOM 2007*, 537-541.

Price, Patti J.; Ostendorf, Mari; Shattuck-Hufnagel, Stefanie and Cynthia Fong. 1991. The use of prosody in syntactic disambiguation. *Journal of the Acoustic Society of America*, 90(6), 2956-2970.

R Core Team. 2022. R: A language and environment for statistical computing. https://www.R-project.org/

Scott, Donia R. 1982. Duration as a cue to the perception of a phrase boundary. *The Journal of the Acoustical Society of America*, 71(4), 996-1007. https://doi.org/10.1121/1.387581

Searle, John R. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press. https://doi.org/10.1017/CBO9781139173438

Searle, John R. 1979. *Expression and Meaning: Studies in the Theory of Speech Acts*. Cambridge University Press. https://doi.org/10.1017/CBO9780511609213

Šturm, Pavel and Aleš Bičan. 2021. *Slabika a její hranice v češtině*. Karolinum.

Uhmann, Susanne. 1992. Contextualizing Relevance: On Some Forms and Functions of Speech Rate Changes in Everyday Conversation. In P. Auer and A. Di Luzio (eds.), *Pragmatics & Beyond New Series*, 297-336. John Benjamins Publishing Company. https://doi.org/10.1075/pbns.22.19uhm

Volín, Jan. 2019. The meaning in language generally and in question-word questions particularly: A study in speech prosody. In T. Hoskovec (ed.), *Expérience et Avenir du Structuralisme. Travaux du Cercle linguistique de Prague*, 123-141. Kanina.

Wickham, Hadley; Averick, Mara; Bryan, Jennifer; Chang, Winston; McGowan, Lucy D'Agostino; François, Romain; Grolemund, Garrett; Hayes, Alex; Henry, Lionel; Hester, Jim; Kuhn, Max; Pedersen, Thomas Lin; Miller, Evan; Bache, Stephan Milton; Müller, Kirill; Ooms, Jeroen; Robinson, David; Seidel, Dana Paige; Spinu, Vitalie; Takahashi, Kohske; Vaughan, Davis; Wilke, Claus; Woo, Kara and Hiroaki Yutani. 2019. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686-1691. https://doi.org/10.21105/joss.01686

Yang, Xiaohong; Shen, Xiangrong; Li, Weijun and Yufang Yang. 2014. How listeners weight acoustic cues to intonational phrase boundaries. *PLoS ONE*, 9(7), e102166. https://doi.org/10.1371/journal.pone.0102166

Zhang, Xinting. 2012. *A Comparison of Cue-weighting in the Perception of Prosodic Phrase Boundaries in English and Chinese*. Ph.D. dissertation. University of Michigan.