2017 ACTA UNIVERSITATIS CAROLINAE PAG. 79–94
PHILOLOGICA 3 / PHONETICA PRAGENSIA

APPEAL AND DISREPUTE OF THE SO-CALLED GLOBAL RHYTHM METRICS

JAN VOLÍN

ABSTRACT

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

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

1. Introduction

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

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

https://doi.org/10.14712/24646830.2017.35

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

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

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

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

Speech clearly cannot afford repeating primitive patterns. Monotony or repetitiveness of simple patterns would constrain its readiness to express a variety of meanings, which is its most precious attribute. Research in rhythm of speech should focus on describing the flow of particular configurations in a given language rather than

on seeking simple primitive regularity. Separating the concept of meter from rhythm might help to avoid the search for a new term to replace the word rhythm in speech sciences.

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

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

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

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

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

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

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

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

2. Method

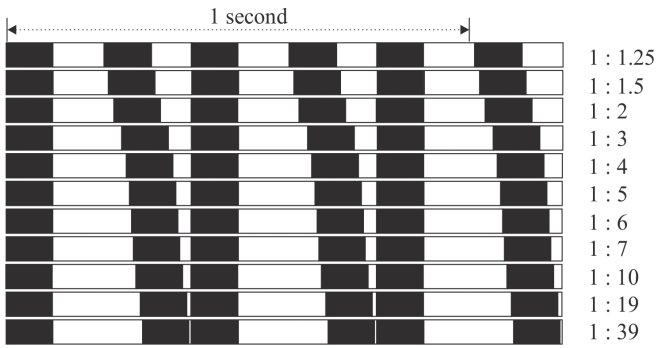
2.1 Simulations

The core (or referential) artificial material was set to emulate the articulation rate of 5 syll/s which is a natural and comfortable tempo for most humans (a faster and slower version was also produced). Durations of consonants were kept constant for the sake of clarity, but the behaviour of metrics on vowels would manifest on any units: the metrics are 'blind' to what they are supposed to measure. Thus, if the examined units were syllables or stress-groups, as long as the ratios are preserved, the metrics would produce iden-

tical results. The intention at this point, however, was to produce material that is based on realistic attributes of speech and, at the same time, is easy to conceptualize.

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

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



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

2.2 Natural material

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

The recordings were divided into breath-groups, i.e., stretches of speech between the intakes of breath. As the speakers were professionals who prepared their reading beforehand, the breath-group boundaries coincided with major syntactic breaks. Phone boundaries were first estimated (forced aligned) by the Prague Labeller (Volín, Skarnitzl & Pollák, 2005; Pollák, Volín, Skarnitzl, 2007) and then manually corrected for various imprecisions. Altogether, slightly over 41.000 consonantal and vocalic boundaries were

checked for further processing. The breath-groups were processed individually, but the arithmetic means calculated later were weighted by their duration. Hence, the shorter units contributed to the overall mean less than longer units.

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

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

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

2.3 Metrics

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

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

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

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

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

The modification of the original formula was suggested by Gibbon and Gut (2001) and it does not change the patterns found in the results. As it is more convenient, it will be used in the current study. However, if a comparison of results is required, the values obtained using the older formula on the left must be halved. (To avoid confusion, Gibbon

and Gut proposed a different name for their adjusted metric. They wanted to call it the Rhythm Ratio. For the reasons stated in the Introduction, this proposal will be ignored.)

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

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

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

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

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

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

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

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

2.4 Procedure

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

Apart from raw measurements in the speech material as a whole, the exclusion of phrase-final portions was carried out to establish its influence on the results. Phrase final

lengthening is a quasi-universal prosodic feature that poses a potential problem to the calculation of the metrics. Its domain might be unstable (anything from the last phone to the last stress-group), but, more importantly, its scale varies hugely. We thus repeated all measurements on the material will all phrase-final words excluded.

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

3. Results

3.1 Artificial Data

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

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

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

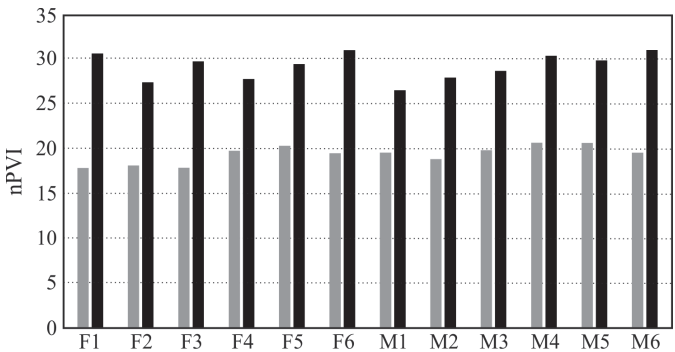
An important thing to note is that *nPVI* is not increasing linearly with the increase of the ratios. That is not surprising from the mathematical point of view, but when people conceptualize their results they should be aware of this. Another mathematically trivial fact is that *ΔS* and *VarcoS* are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

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

3.2 Natural Data

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

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

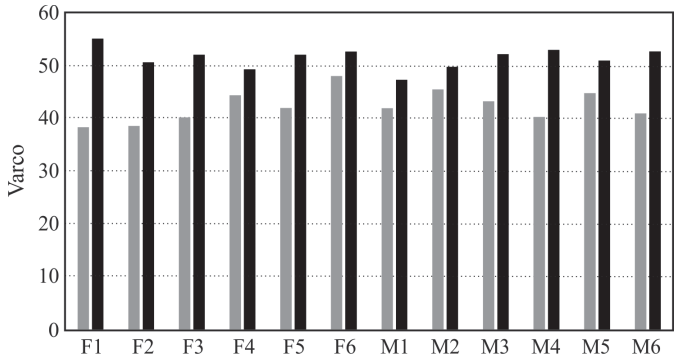


It can be observed that consonantal pairwise variation is greater than the vocalic one. Consonant clustering is relatively common in Czech. The existence of phonological length in the vowel system does not seem to have any particularly profound influence on variation in vowel durations. In comparison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

The values of *Varco* seem to be less compact than the previous metric: they ranged between 38.4 and 48.1 for vowels and between 47.4 to 55.2 for consonants (see Fig-

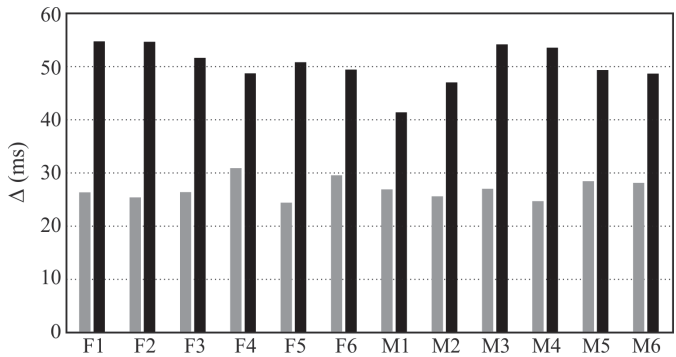
ure 3). Their insignificant mutual correlation also confirmed independence of vocalic and consonantal variation. However, the correlation between *nPVI-V* and *Varco-V* was established at $r = 0.55$ ($p < 0.001$) and for consonants (*nPVI-C* × *Varco-C*) even $r = 0.76$ ($p < 0.001$). Although this is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

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



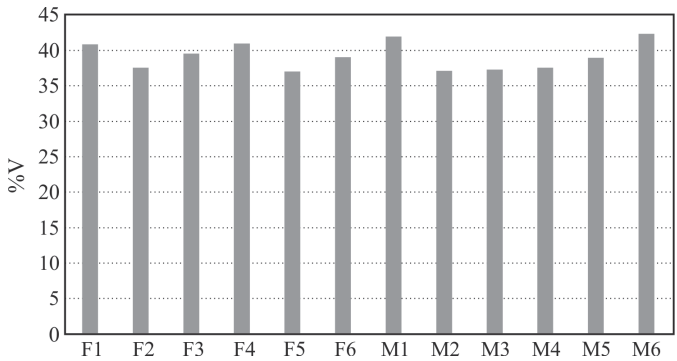
Non-normalized counterparts of *Varco* measures are standard deviations from mean durations labelled Δ after Ramus et al. (1999). They are displayed in Figure 4. The Δ*V* values range from 24.4 to 30.9 ms, whereas the Δ*C* values from 41.1 to 54.7 ms. The correlation coefficient for Δ*V* against *VarcoV*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

Figure 4. Values of Δ*V* and Δ*C* (standard deviations from mean durations) for individual speakers. Black columns represent consonants, grey columns represent vowels.



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

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



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

3.2.1 Local changes in tempo

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

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

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

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

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

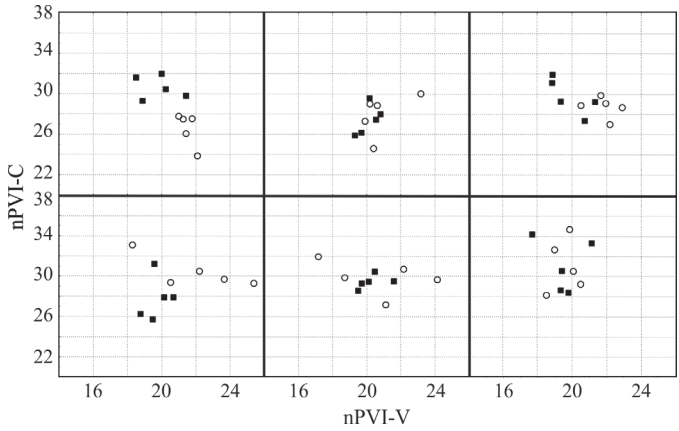
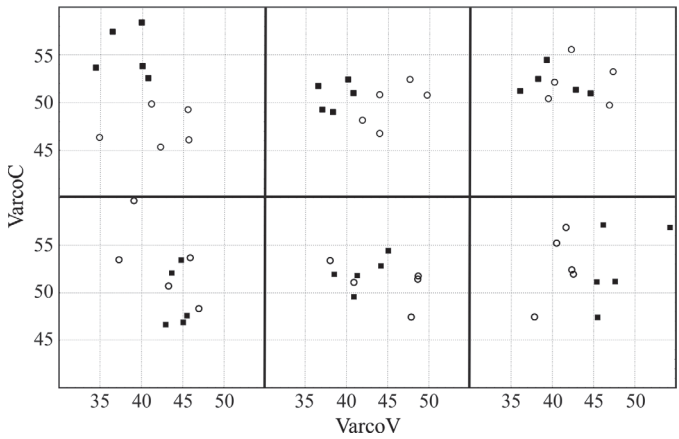


Figure 7. Scatterplots of *Varco* values of randomly paired speakers, calculated for random fifths of their spoken texts. Male speakers are represented by empty circles, female speakers by filled squares.



3.2.2 Intraspeaker variation

The results depicted in the Figures 2 to 5 above show that the interspeaker variation is not very high, but we felt it necessary to establish the magnitude of intraspeaker variations as well. In Figures 6 and 7 the speakers were randomly paired (by alphabetic cues

from their surnames). In each of the scatterplots there are nPVI or Varco values of a male and a female speaker. As explained in the Method, the five values represent five random fifths of their breath-groups, i.e., about 100 words of the spoken text.

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

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

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

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

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

3.2.3 Czech versus English

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

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

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

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

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

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

4. Discussion

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

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

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

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

a passage"). Current research community will quite certainly agree that such practice may bring mistrust to the research field and should, therefore, become a thing of the past.

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

ACKNOWLEDGEMENT

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

REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Kohler, K. (2009). Rhythm in speech and language. *Phonetica*, 66, 29–45.
Loukina, A., Kochanski, G., Rosner, B., Keane, E. & Shih, Ch. (2011). Rhythm measures and dimensions of durational variation in speech. *Journal of Acoustical Society of America*, 129/5, 3258–3270.
Low, E. L., Grabe, E. & Nolan, F. (2000). Quantitative characterisations of speech rhythm: Syllable timing in Singapore English. *Language and Speech*, 43, 377–401.
Low, E. L. & Grabe, E. (1995). Prosodic patterns in Singapore English. In: *Proceedings of 13th International Congress of Phonetic Sciences*, 636–639.
Mariano, P. & Romano, A. (2011). Rhythm metrics for 21 languages. In: *Proceedings of 17th ICPhS*, 1318–1321.
Miller, G. R. & Hewgill, M. A. (1964). The effect of variations in non-fluency on audience ratings of source credibility. *Quarterly Journal of Speech*, 50/1, 36–44.
Nolan, F. & Jeon, H.-S. (2014). Speech rhythm: a metaphor? *Philosophical Transactions of the Royal Society B*, 1–11.
O'Rourke, E. (2008). Speech rhythm variation in dialects of Spanish: applying the pairwise variability index and variation coefficients to Peruvian Spanish. In: *Speech Prosody 2008*, 431–434.
Pollák P., Volín, J. & Skarnitzl, R. (2007). HMM-based phonetic segmentation in Praat environment. In: *Proceedings of 12th Intern. Conf. Speech & Computer – SPECOM 2007*, 537–541, Moscow: MSLU.
Quené, H. & Port, R. F. (2005). Effects of timing regularity and metrical expectancy on spoken-word perception. *Phonetica*, 62, 1–13.
Ramus, F., Nespor, M. & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, 73, 265–292.
Skarnitzl, R. & Volín, J. (2012). Referenční hodnoty vokalických formantů pro mladé dospělé mluvčí standardní češtiny [Reference values of vowel formants for young adult speakers of Standard Czech]. *Akustické listy*, 18/1, 7–11.
Slówik, O. & Volín, J. (in print). Acoustic correlates of temporal structure in North-Vietnamese English. In: J. Volín & R. Skarnitzl. (Eds.), *The Pronunciation of English by Speakers of Other Languages*. Newcastle: Cambridge Scholars Publishing.
Volín, J. & Churaňová, E. (2010). Probabilities of consonantal sequences in continuous Czech texts. *AUC – Philologica 1, Phonetica Pragensia XII*, 49–62.
Volín, J. & Skarnitzl, R. (2007). Temporal downtrends in Czech read speech. In: *Proceedings of 8th Interspeech*, 442–445.
Volín, J., Skarnitzl, R. & Pollák, P. (2005). Confronting HMM-based phone labelling with human evaluation of speech production. In: *Proceedings of Interspeech 2005*, 1541–1544.
Wagner, P. & Dellwo, V. (2004). Introducing YARD and re-introducing isochrony to rhythm research. In: *Proceedings of Speech Prosody 2004*.
White, L., Mattys, S., Series, L. & Gage, S. (2007). Rhythm metrics predict rhythmic discrimination. In: *Proceedings of 16th ICPhS*, 1009–1012.

RESUMÉ

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

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