RHYTHM METRICS FOR SPEAKER IDENTIFICATION IN CZECH

LENKA WEINGARTOVÁ

ABSTRACT

This paper investigates usefulness of the global rhythm metrics (such as %V, ΔC, PVI, Varco, etc.), introduced by Ramus et al. (1999), Grabe and Low (2002) and others, for speaker identification purposes. In our sample of three Czech female speakers, these features failed to capture the inter-speaker differences satisfactorily. They are furthermore put in comparison with a local articulation rate (LAR) measure (Volín, 2009), which is shown to be capable of capturing phrase-final lengthening differences between all speakers.

Key words: phrase-final lengthening, rhythm metrics, timing, speaker identification

1. Introduction

The rhythm of speech is a complex phenomenon that has yet to be fully understood and satisfactorily described. There is a number of different approaches to this issue – among the most influential are coupled neural oscillators modelling (pursued in Šimko and Cummins, 2010; Large et al., 2010; etc.), articulatory gesture modelling (e.g. Saltzman and Munhall, 1989; Saltzman et al., 2008, or Gafos and Goldstein, 2012) and the somewhat simpler rhythmic or temporal metrics (Ramus et al., 1999; Low et al., 2000; Grabe and Low, 2002; Asu and Nolan, 2006; Dellwo, 2006, or Arvaniti, 2009 and 2012).

The modelling approaches will not be pursued here for a number of reasons. Their top-down attitude sooner or later encounters the problem of data fitting, where unwieldy natural speech clashes with the elegance of the model. Building such models requires a priori knowledge or hypotheses – and since there is not enough data on Czech to formulate such hypotheses, it seems reasonable to employ an inverse approach and first look into the available material for possible generalizations.

The rhythm metrics are much more tempting – they are easy to extract from the material and to employ for a new language. They evolved from repeated attempts to quantify the infamous stress-timing vs syllable-timing language distinction (first formulated by Pike, 1945) and have been used – with varying rates of success – for a number of different problems since then (dialect discrimination, speech impairment quantification, speaker recognition, etc.).

Before we turn to the details, it should be pointed out that none of these approaches actually describe speech rhythm as a whole. In some way or another, all of them employ information about the timing of speech events only. But there is much more to rhythm than just temporal structuring and duration of speech sounds: we have to consider prosodic information, such as stress or F0 changes, and its influence on perception of duration (see for example Cumming, 2001), issues of neurological processing (e.g. the emergence of subjective rhythm, as in Fraisse, 1982), etc. The number of factors influencing rhythm of speech and its production and perception is indeed vast.

Therefore, when attempting to describe or objectively quantify speech rhythm, there are sacrifices to be made. As there is no holistic theory of rhythm to date, the logical thing is to come at this complex problem from simpler points of view – in this case, we examine timing, and more specifically, the individuality of timing of different speakers.

The rhythm metrics (and the word "rhythm" being used extremely reluctantly here and purely out of tradition) offer us a way to globally quantify certain aspects of speech segments duration. Although they have originally been used for language classification, many researchers have reported certain degree of speaker-dependence (e.g. Ramus, 2002; Asu and Nolan, 2006; White and Mattys, 2007; Dellwo and Koreman, 2008, or Arvaniti, 2009) at least for some of them. The metrics describe the variability of vocalic and consonantal intervals, which can be expected to vary from speaker to speaker, if they treat the durations of segments differently.

The main question posed in the first part of the experiment is to what extent are these metrics dependent on the individuality of the speaker and whether they are able to differentiate between speakers of the same language.

Nevertheless, caution is needed in interpreting these metrics linguistically. A lot of factors can play a role in the variability of vocalic and consonantal intervals. For example, it has been shown on experiments with impaired speech (Lowit, 2012) that very different underlying causes can result in similar scores of rhythm metrics.

The convenient advantage and at the same time the greatest constraint of these metrics is their globality. They describe global temporal characteristics of a given utterance, although the timing differences of speakers might occur only locally. In the second part of the experiment, we will therefore look at one particular phenomenon that may exert a considerable influence on the duration of segments – phrase-final lengthening (see also Arvaniti, 2009). If the speakers were to differ in the extent of their final lengthening, it is not unreasonable to suppose that it will show in differences between the timing metrics.

However, if we want to look directly at local timing changes, the global parameters do not seem to be a sensible choice. To use a global metric on a local phenomenon means to severely reduce the information value of the metric which needs some minimal amount of data to be reliable. A local temporal metric called LAR (Local Articulation Rate, developed by Volín, 2009) was therefore applied to distinguish speakers according to their phrase-final lengthening. The localized approach to speech timing as opposed to the global metrics has many advantages and will be further discussed in the final section of this paper.

2. Method

For both experiments, six recordings of three female professional speakers (two recordings from each speaker) from the Prague Phonetic Corpus (Skarnitzl, 2010) were used. The average total length of utterances from one speaker was 6.5 minutes and 934 words. The texts differed in content but not in style (newsreading).

The recordings were segmented to breath-groups, automatically labelled with the Prague Labeller tool (Pollák et al., 2007) in Praat (Boersma and Weenink, 2012) and manually corrected by the author using guidelines from Machač and Skarnitzl (2009). Vocalic and consonantal intervals were labelled using a Praat script, syllabic consonants were treated as vowels for this purpose. Consonantal groups across word boundaries were treated as one C interval. All intervals containing pauses or dysfluencies were disregarded.

The following parameters were extracted:

- %V: the proportion of vocalic intervals within a breath-group,
  i.e., %V = $(d_{vt}/d_t) \times 100$,
  where d_{vt} is the total duration of vocalic intervals within a breath-group and d_t the total duration of the breath-group
- ΔV/ΔC: the standard deviation of the duration of vocalic/consonantal intervals within a breath-group,

  i.e., $\Delta V = \sqrt{\dfrac{\Sigma(d_v\text{-}d_{vavg})^2}{n\text{-}1}}$,

  where d_v is the duration of one vocalic interval and d_{vavg} is the mean duration of vocalic intervals in the breath-group; similarly for ΔC
- VarcoV/VarcoC: ΔV or ΔC normalised by average vocalic/consonantal duration within a breath-group,
  i.e., VarcoV = $\Delta V/d_{vavg}$,
  similarly for VarcoC
- rPVI-V/rPVI-C: raw Pairwise Variability Indices for vocalic/consonantal intervals,

  i.e., $rPVI = [\Sigma_{k=1}^{n-1}|d_k - d_{k+1}| \,/(\mathrm{n}-1)]$,

  where d_k is the duration of k-th interval and n the number of respective intervals
- nPVI-V/nPVI-C: normalised Pairwise Variability Indices for vocalic/consonantal intervals,

  i.e., $nPVI = 100 \times [\Sigma_{k=1}^{n-1}|\frac{d_k - d_{k+1}}{(d_k + d_{k+1})/2}| \,/(\mathrm{n}-1)]$,

- LAR: local articulation rate values; see below

For the extraction of LAR values, midpoints of each vocalic interval had to be labelled – the inverse value of the distance of two successive midpoints ($dur_{pk\text{-}pk}$) then constitutes one LAR value:

$$LAR = \frac{1}{dur_{pk-pk}}$$

The inverse value has the effect of converting distance measure to articulation rate measure, therefore the higher LAR value, the higher articulation rate in syllables per seconds.

For analysing the phrase-final lengthening, final sections of breath-groups in the length of six syllables were chosen. Only those breath-groups were selected which exhibited a falling nuclear tone and whose last intonation phrase was longer than six syllables (a prosodic boundary within this section would interfere with the final lengthening and the breath-groups would not be comparable). The last six LAR values were measured and smoothed using the 3-point moving average method. A linear regression coefficient of each of the breath-group endings was computed using the method of least squares, representing the overall gradient of rising or falling of the LAR values. The moving average smoothing served to diminish the drops in LAR caused mainly by consonantal clusters.

As the texts of each recording did not have identical content, a control measure was employed to quantify the effect of syllable structure on all the metrics. A simple proportion of the number of individual consonants and vowels for each breath-group (C/V proportion) and its correlation with the measures was computed. Any breath-groups showing this proportion higher or lower than 2 standard deviations from the overall average were discarded (this was mainly the case of breath-groups containing a foreign proper name).

Another investigated factor was the percentual representation of phonologically long vowels in the text, which also could lead to differences in the metrics without showing any speaker idiosyncrasies.

After preliminary analyses, some additional post-hoc modifications were made: 5% of the longest and shortest breath-groups in terms of total duration of the articulation were removed (such as the greetings at the beginning of each recording) as they seemed to introduce additional variability which obscured the rest of the data.

All extracted measures were compared pairwise (each speaker with each other and the two recordings of one speaker) using t-tests for uncorrelated samples with homogeneous variance; correlations were computed using Pearson's or Spearman's coefficients.

3. Results

3.1 Global rhythm metrics

Although the metrics and the C/V proportions were to some extent correlated (which was to be expected), there was no statistically significant difference between the C/V proportions of the data from each speaker. This led us to conclude that the recordings are phonotactically representative and should not introduce any artifacts into the data.

Furthermore, there was no significant correlation between the percentual proportion of long vowels and the values of the vocalic measures, so the presence of long vowels does not affect them to a great extent.

The results of the between- and intra-speaker analyses for all nine parameters is shown in Table 1 below.

Table 1. Statistical significance of the results of pairwise t-tests comparing global rhythmic values for each breath-group. The first three columns show inter-speaker differences, the second three intra-speaker. A difference of $p < 0.05$ is marked by asterisk.

Speaker/Measure	F1/F2	F2/F3	F1/F3	F1a/F1b	F2a/F2b	F3a/F3b
%V		*	*			
ΔV	*	*			*	*
VarcoV					*	
rPVI-V	*	*				
nPVI-V						
ΔC						
VarcoC	*					
rPVI-C						*
nPVI-C	*		*			

In general, the vocalic measures performed better than consonantal, but none of the timing metrics was able to discriminate all three speakers from each other. In addition, some of them (especially ΔV) even discriminated between two recordings of the same speaker, which is unwanted in this case. VarcoV, nPVI-V, ΔC and rPVI-C did not show any inter-speaker differences at all.

We also tried to improve the results by combining the more promising measures (%V, rPVI-V and nPVI-C) with each other in two-dimensional scatterplots, but this still did not reveal any significant trends. An example of combining %V with nPVI-C is shown in the figure below.

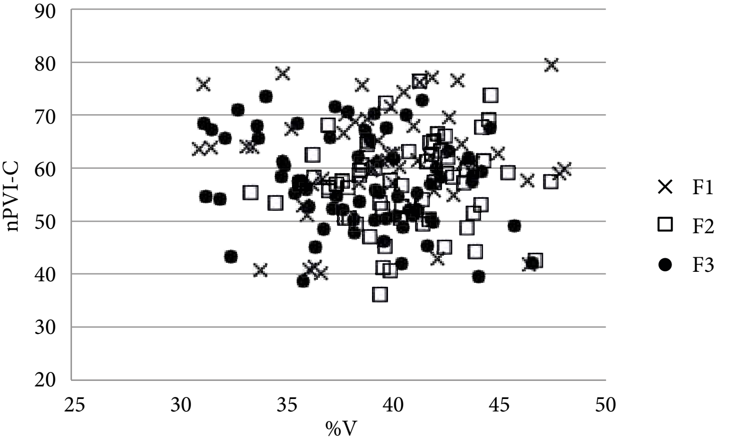

Figure 1. Values of %V and nPVI-C for all three speakers. Each point represents the value of one breath-group.

3.2 LAR values

Turning now to the LAR regression values representing the rate of final lengthening, the picture turns out to be more encouraging.

In Figures 2a–b, there are two examples of what a smoothed LAR contour looks like. Both represent the last six syllables of a breath-group ending with a falling nuclear tone. There are only 4 LAR values, since the 3-point smoothing averages out two of the values. Figure 2a shows a distinct final lengthening (the lower LAR values the bigger distance between two syllable nuclei), Figure 2b on the other hand shows acceleration. Both tendencies are captured by the linear regression coefficient – a negative value of the coefficient equals slowing down, a positive number indicates speeding up.

Figure 3 shows overall average values and standard deviations of LAR linear regression coefficients for the three speakers, in Figure 4 they are broken up to individual recordings.

It is evident from Figure 3 that all speakers in general use lengthening of vowels at the end of utterances – the averages of the coefficients are negative. Speaker F1 lengthens almost exclusively all of her breath-group endings, speaker F2 approximately half of the cases. This is exactly the kind of speaker-specific behaviour we have been looking for. Although there are considerable intra-speaker variations, the speakers within each pair are significantly different from each other – the results of t-tests are shown in Table 2 below.

Moreover, the regression coefficient of LAR did not differentiate the two recordings from one speaker (see Figure 4), which is favourable in speaker identification cases, where intra-speaker variability can be a non-trivial problem.

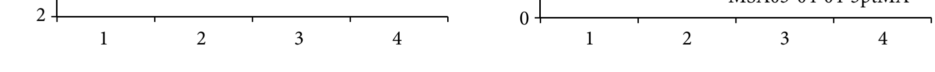

Figure 2a–b. Individual LAR contours for the last six syllables of two selected utterances. LAR values are smoothed using a 3-point moving average. Straight line represents the linear regression line, with its coefficient shown above it.

Table 2. Results of t-tests comparing all pairs of speakers.

F1/F2	F2/F3	F1/F3
$t\,(65) = 4.16$	$t\,(72) = 2.14$	$t\,(61) = 2.3$
$p < 0.05$	$p < 0.05$	$p < 0.05$

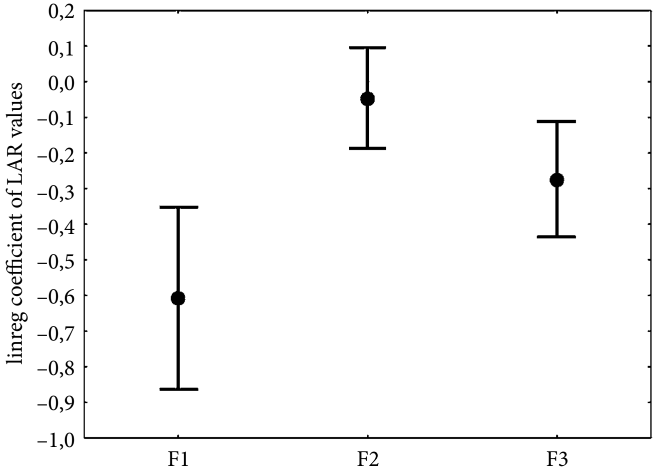

Figure 3. Mean values and standard deviations of linear regression coefficients of the 3-point moving average LAR values for all three speakers. Whiskers denote 0.95 confidence interval.

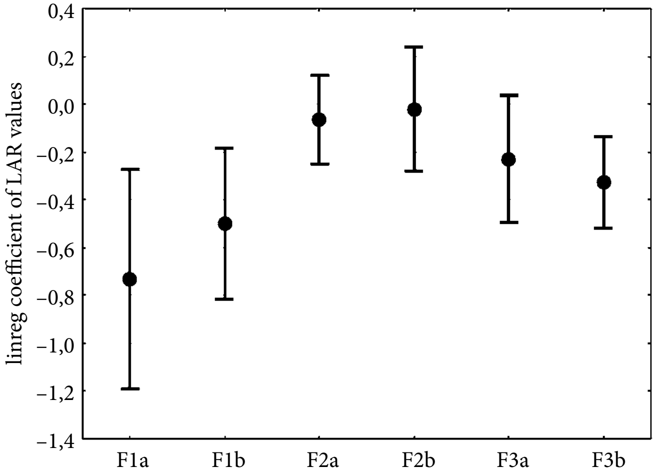

Figure 4. Mean values and standard deviations of linear regression coefficients of the 3-point moving average LAR values for all six recordings. Whiskers denote 0.95 confidence interval.

4. Conclusion and discussion

Considering the results, a simple conclusion can be drawn: the local LAR measure is superior in capturing inter-speaker differences in our material to the global timing metrics. Although the sample of speakers is very small to generalize, we believe that looking at local temporal changes may be more useful than trying to cluster all the information about timing in an utterance into one number. It seems that the global measures are indeed more apt to distinguish languages from one another, as the local speaker-specific differences integrate into fairly similar results.

As a sidenote, it is interesting to point out that the average values of %V for all three speakers cluster around 40, which would place Czech next to German with very similar results (Arvaniti, 2012). Comparing the results of global metrics with other languages as reported in the literature, it is remarkable that according to our (albeit sparse) data, Czech has very low values of all other vocalic measures, which is surprising, since one would expect a language with vowel quantity contrast to have higher variability in vowels. This suggests that the timing metrics are unable to capture the contrast (for similar results with Hungarian, Finnish and Turkish see also Papp, 2012) and, furthermore, that Czech short and long vowels do not differ that significantly in duration – this notion is supported by Skarnitzl and Volín (2012) who found that the proportion of long to short vowel durations in Czech ranges only from 1.29 to 1.79 and is even smaller in nonfinal positions.

It could be also interesting to turn the attention to the question of perceptual saliency of the global metrics. As Dellwo pointed out (Dellwo et al., 2012), it could well be the case that utterances with very different temporal values might not be that different perceptually. If so, then the metrics are not able to tell us anything about how the human brain distinguishes speakers. But, on the other hand, if people are not aware of such information, it suggests that it cannot be intentionally manipulated and therefore could be useful as a forensic feature.

In further research we plan to continue investigating local measures and look directly at temporal trajectories of utterances and their speaker-specific characteristics. The aim is to find language-specific timing patterns for Czech and see how speakers express their individual speech habits in comparison with the average temporal behaviour. Some preliminary results can be found in Volín and Weingartová (2012).

However, the intra-speaker variability remains an open question. Examination of more recordings from one speaker could shed light on robustness of the local timing metrics and on consistency of the timing patterns produced by one speaker.

Another logical step will be to move forward from read speech to more spontaneous speech, to test whether the hypotheses hold. A great advantage of measuring timing, especially when we have forensic uses in mind, is its robustness to noise and to intentional disguise of the speaker. We therefore believe that, once better understood, temporal patterns of speech may become a very useful tool in forensic speaker identification.

Also, a combined approach that extends the scope to other prosodic domains (such as F0 contours or speech energy dynamics, as it is done by Adami et al., 2007) and their timing relationships could considerably increase the success rate of speaker recognition.

ACKNOWLEDGEMENTS

This research was supported by a project of the Czech Science Foundation (GAČR 406/12/0298) and by the Programme of Scientific Areas Development at Charles University in Prague (PRVOUK), subsection 10 – Linguistics: Social Group Variation. I would also like to express most sincere thanks to my supervisor, Jan Volín, for his help and guidance.

REFERENCES

Adami, A. G., Mihaescu, R., Reynolds, D. A. & Godfrey, J. J. (2007). Modeling prosodic differences for speaker recognition, Speech Communication, 49/4, pp. 277–291.

Arvaniti, A. (2009). Rhythm, timing and the timing of rhythm, Phonetica, 66, pp. 46–63.

Arvaniti, A. (2012). The usefulness of metrics in the quantification of speech rhythm. Journal of Phonetics, 40, pp. 351–373.

Asu, E. L. & Nolan, F. (2006). Estonian and English rhythm: A two-dimensional quantification based on syllables and feet. In: Proceedings of Speech Prosody 2006, Dresden, Germany.

Boersma, P. & Weenink, D. (2012). Praat: doing phonetics by computer [Computer program]. Version 5.2.19, retrieved on March 20, 2012 from <http://www.praat.org>.

Cumming, R. (2001). The effect of dynamic fundamental frequency on the perception of duration. Journal of Phonetics, 39/3, pp. 375–387.

Dellwo, V. (2006). Rhythm and Speech Rate: A Variation Coefficient for C. In: P. Karnowski & I. Szigeti (Eds.), Language and Language processing. Frankfurt am Main: Peter Lang, pp. 231–241.

Dellwo, V., Kolly, M. & Leemann, A. (2012). Speaker identification based on temporal information: A forensic phonetic study of speech rhythm and timing in the Zurich variety of Swiss German. In: Proceedings of IAFPA 2012. Santander: IAFPA.

Dellwo, V. & Koreman, J. (2008). How speaker idiosyncratic is measurable speech rhythm? In: Proceedings of IAFPA 2008. Lausanne: IAFPA.

Fraisse, P. (1982). Rhythm and tempo. In: D. Deutsch (Ed.), The Psychology of Music. New York: Academic Press, pp. 149–180.

Gafos, D. & Goldstein, L. (2012). Articulatory representation and organization. In: A. Cohn, M. Huffman & C. Fougéron (Eds.), Handbook of Laboratory Phonology. Oxford: Oxford University Press, pp. 220–231.

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

Large, E. W., Almonte, F. & Velasco, M. (2010). A canonical model for gradient frequency neural networks. Physica D, 239, pp. 905–911.

Low, E. L., Grabe, E. & Nolan, F. (2000). Quantitative Characterizations of Speech Rhythm: Syllable--Timing in Singapore English. Language & Speech, 43/4, pp. 377–401.

Lowit, A. (2012). Application of rhythm metrics in disordered speech: What should we measure and what do they really tell us? In: Proceedings of Perspectives on Rhythm and Timing. Glasgow: UG, p. 36.

Machač, P. & Skarnitzl, R. (2009). Principles of Phonetic Segmentation. Praha: Epocha.

Papp, V. (2012). Rhythmic typology of three languages with vowel harmony and quantitative contrast. In: Proceedings of Perspectives on Rhythm and Timing. Glasgow: UG, p. 48.

Pike, K. L. (1945). The intonation of American English. Ann Arbor: University of Michigan Press.

Pollák, P., Volín, J. & Skarnitzl, R. (2007). HMM-Based Phonetic Segmentation in Praat Environment. Proceedings of the XIIth International Conference "Speech and computer – SPECOM 2007". Moscow: MSLU, pp. 537–541.

Ramus, F., Nespor, M. & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. Cognition, 73/3, pp. 265–292.

Ramus, F. (2002). Acoustic correlates of linguistic rhythm: Perspectives. In: Proceedings of Speech Prosody 2002. Aix-en-Provence: ISCA, pp. 115–120.

Saltzman, E. L. & Munhall, K. (1989). A dynamical approach to gestural patterning in speech production. Ecological Psychology, 1/4, pp. 333–382.

Saltzman, E., Nam, H., Krivokapic, J. & Goldstein, L. (2008). A task-dynamic toolkit for modeling the effects of prosodic structure on articulation. In: Proceedings of Speech Prosody 2008, Campinas.

Skarnitzl, R. (2010). Prague Phonetic Corpus: status report. AUC Philologica 1/2009, Phonetica Pragensia, XII, pp. 65–67.

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

Šimko, J. & Cummins, F. (2010). Embodied task dynamics. Psychological Review, 117/4, pp. 1229–1246.

Volín, J. (2009). Metric warping in Czech newsreading. In: R. Vích (Ed.), Speech Processing – 19th Czech-German Workshop. Praha: ÚFE AVČR, pp. 52–55.

Volín, J. & Weingartová, L. (2012). Idiosyncrasies in local articulation rate trajectories in Czech. In: Proceedings of Perspectives on Rhythm and Timing. Glasgow: UG, p. 67.

White, L. & Mattys, S. L. (2007). Calibrating rhythm: First language and second language studies. Journal of Phonetics, 35, pp. 501–522.

RYTMICKÉ UKAZATELE PRO ROZPOZNÁVÁNÍ MLUVČÍHO V ČEŠTINĚ

Resumé

Článek se zabývá využitím globálních rytmických ukazatelů pro identifikaci mluvčího. Tyto metriky (%V, ΔC, PVI, Varco, atd.) zavedené ve studiích Ramuse et al. (1999), Grabové a Lowové (2002) a dalších nicméně ve vzorku tří českých mluvčích ženského pohlaví nedokázaly uspokojivě zachytit rozdíly. Jsou tedy dále srovnávány s lokálním popisem tempa (s pomocí proměnné LAR, viz Volín, 2009), zaměřeným na závěrové zpomalování na koncích promluvových úseků. Tento lokální popis byl již schopen zachytit rozdíly mezi mluvčími v míře závěrového zpomalování.