INTERSPEECH 2009 BRIGHTON

The Dynamic Dimension of the Global Speech-Rhythm Attributes

Jan Volín [1], *Petr Pollák* [2]

[1] Institute of Phonetics, Charles University in Prague, Czech Republic
[2] Faculty of Electrical Engineering, Czech Technical University in Prague, Czech Republic
jan.volin@ff.cuni.cz, pollak@fel.cvut.cz

Abstract

Recent years have revealed that certain global attributes of speech rhythm can be quite successfully captured with respect to consonantal and vocalic intervals in spoken texts. One of the problems of this approach lies in complex syllabic structures. Unless we make an a-priori phonological decision, sonorous consonants may contribute to either vocalic or consonantal part of the speech signal in post-initial and pre-final positions of syllabic onsets and codas. A procedure is offered to avoid phonological dilemmas together with tedious manual work. The method is tested on continuous Czech and English texts read out by several professionals.

Index Terms: speech rhythm metrics, rhythm class

1. Introduction

Despite the fact that common variations in speech rhythm do not affect the lexical meanings of words in any direct manner and are therefore not considered phonological in the traditional sense of the term, the phenomenon of prominence alternation and its effects on listeners have always been regarded as interesting and attractive to researchers. With the exception of misplaced word stresses, the temporal attributes of lighter and heavier position exchanges in the linear progression of utterances seem to influence the impression that the speaker makes on the listener, but not the informational contents of a linguistic unit per se.

However, this effect is not a trivial one and neither are the potential consequences of the deformations of the modal rhythmic forms. Buxton [1] offered an overview of studies which showed that alterations of natural rhythm flow led to longer reaction times in monitoring experiments. Listeners seem to exert some extra mental effort when they are asked to spot a word or a syllable in the speech continuum which deviates from predictable rhythmic pattern.

The neurophysiological foundation of this additional processing load is perhaps best explained by models offered in [2]. The concept of the neural resonance mechanism, which is generalizable beyond the domain of speech, builds on converging evidence that to recognize an object in a given context a dedicated neural assembly in the brain has to perform an act of resonance. It is a series of synchronous activations of neurons which occurs when the expectational neural representations (based on experience and the concurrent context analysis) meet with the input neural representations (based on the properties of the incoming signal). The timing of the two main streams of neural activity is crucial. If the internal, expectational impulses meet the incoming representations in reasonable synchrony, the process of perception is smooth and relatively effortless. If, however, the timing of the incoming events is unpredictable, the relevant neurons have to repeat their activities or draw on other resources until the state of resonance is achieved. Extending the argument of [2] then, speech perception is satisfactorily efficient as long as the temporal structure of the incoming signal is predictable.

If the possible outcome of faulty temporal structures can cause discomfort, albeit unconscious, then it is clearly worth studying. Current research offers two major categories of approaches to the speech rhythm challenge. One of them focuses on determining positions in the speech signal of the occurrence of rhythmic beats from both production ([3], [4]) and perception ([5], [6], [7], [8]) perspective. The other concerns the rhythm typology of languages and tries to establish reliable global attributes of the speech continuum which would correlate with the listeners' intuitions about stress-timing, syllable-timing or mora-timing.

The latter revolves around two fundamental concepts introduced in [9] and [10]. Although these proposals of global rhythm metrics were subjected to justified criticism, e.g., [11], and motivated attempts to correct or improve them, e.g., [12] [13], they inspired a number of studies, many of which brought interesting results ([14], [15], [16]).

In spite of their shortcomings, the global rhythmic metrics (see below, Section 2) allow for testing valuable linguistic hypotheses about differences between languages, language accents, speech styles or personal idiosyncrasies. However, their use has, to our best knowledge, always been dependent on manual labeling of the speech signal events. The metrics require relatively precise placement of boundaries between vocalic and consonantal stretches of the speech continuum.

This is objectionable for two reasons. First, manual labeling can be quite tiresome, especially if one wants to work with larger corpora to avoid artifacts peculiar to limited speech samples. Second, it is contradicting one of the original motivations of one of the proposals. The authors in [9] pointed out that newborn babies recognize the rhythmic features of their mother tongue despite the lack of linguistic knowledge about the word or syllable boundaries. They rely solely on distinction between high-energy portions of the speech signal, which represent vowels, and low-energy stretches, which are supposedly consonants. The logic of this argument is agreeable and in languages with simple syllable structure quite realistic. However, to carry the argument a bit further we should remember that just as the newborns possess no knowledge about word/syllable boundaries, they also do not know that high-energy segments called sonorant consonants should not be confused with vowels from the phonological point of view. This dilemma might be quite important in languages like Czech or English which permit complex consonant clusters. For instance, the word *plan* has a sonorant /l/ in the post-initial position, while the word *pulp* has it in the pre-final position. Does this segment contribute to the perceived rhythmic pattern as a consonant or as a vowel? Given that more than 20 % of phones in Czech or English connected speech are sonorant consonants, the matter of their role in rhythm perception should be taken seriously. One of the key questions in our study therefore is, whether speech

rhythm is perceived through some sort of a phonological module and, despite their high energy, sonorants are taken as consonantal elements or whether it is precisely the amount of energy in the speech segment which will determine its role in the speech rhythm patterning.

In Study I we would like to compare the global rhythm metrics of [9] and [10] for Czech and English based on the conventional phonologically motivated manual labeling with those based on energy distributions in the speech signal. The fundamental question in Study I is whether the latter can capture rhythm features in a similar manner as the former.

Study II tests the rhythm metrics based on signal energy distribution on a more extensive sample. This sample is not manually labeled but its purpose is to verify whether the results of Study I reflect peculiar speaking habits of the two randomly chosen speakers or whether they are generalizable for the given speaking style in Czech and English.

2. Method

2.1. Material

Two recordings (1 Czech and 1 English) of a news bulletin broadcast by a national radio station were retrieved and sampled at 16 kHz with 16-bit resolution for Study I. The bulletins were read by male professional newsreaders and comprised about 500 words each (i.e., about 3.5 minutes of speech). Each news bulletin featured 7 different topics.

The manual labeling was carried out by two experienced phoneticians who identified individual phone boundaries with greatest possible care. Their work relied on visual inspection of spectrograms and auditory cues.

Six additional recordings were used in Study II. These were identical in nature with the ones used in Study I, but were read by different male speakers and were not manually labeled since Study II did not compare manual labeling with automatic extraction. Altogether eight news bulletins read by male speakers (4 Czech and 4 English) were employed.

2.2. Procedure

2.2.1. Manually extracted parameters

The exact duration of all consonantal (C) and vocalic (V) intervals was measured. As suggested in [9, 10], no attention was paid to syllable or word boundaries. The four most commonly used rhythm metrics were calculated:

(a) %V = percentage of the V interval durations within the overall duration of speech activity
(b) ΔC = standard deviation of the C interval durations from the mean (symbol adopted from [9])
(c) PVI-V = pairwise variability index for V intervals
(d) PVI-C = pairwise variability index for C intervals

The PVI was calculated as follows (see [12]):

$$PVI = 100 \times \sum_{k=1}^{m=1} \left| \frac{d_k - d_{k+1}}{d_k + d_{k+1}} \right| / (m-1), \quad (1)$$

where d_k and d_{k+1} are durations of two consecutive vocalic or consonantal intervals and m is the number of such intervals in a measured stretch of speech. Our material was divided into physiological breath-groups, i.e., the rhythm metrics were calculated for the stretches of speech between two breaths. Phrase final lengthening was disregarded as it did not change the results of the present study in any noteworthy manner.

2.2.2. Automatically extracted parameters

It is more fitting to call the automatically extracted regions high-energy (HE) and low-energy (LE) intervals rather than vocalic and consonantal intervals since we no longer rely on phonological status of individual phones. As explained above, sonorant consonants now contribute to either the HE or LE measures depending on their specific properties at a given moment. Durations of HE/LE intervals were used in parallel to the previous routine to calculate the following four metrics:

(a) %HE = percentage of the HE interval durations within the overall duration of speech activity
(b) ΔLE = standard deviation of the LE interval durations from the mean (symbol parallel to [9])
(c) PVI-HE = pairwise variability index for HE intervals
(d) PVI-LE = pairwise variability index for LE intervals

The detection of HE/LE intervals was based on simple power analysis commonly used for, e.g., VAD (voice activity detection)[17] or energy-based syllable detection as presented in [18]. It entails the following principal steps: (1) short-time energy computation, (2) comparison of short-time energy with a proper threshold, and (3) possible assessment of outcomes in more frequency bands.

Since the ***short-time energy computation*** is a relatively well-known and frequently used technique, only selected aspects will be described.

- Primary short-time power ($P(i)$) computation is based on block technique computed for frames of length N with 50% overlapping. The index i represents i-th frame.
- Suitable settings of frame length N is important for time resolution of boundary locations. For given N with 50% overlap this resolution is $N/2$. We chose N = 10 ms.
- To eliminate short-time fluctuations, the smoothing of primary power was accomplished by recursive exponential averaging with time constant set to 25 ms:

$$\overline{P}(i) = q \cdot \overline{P}(i-1) + (1-q) \cdot P(i), \quad (2)$$

where q ~ 1 – 4·25/N and N is in ms.

The ***decision threshold*** for HE/LE interval distinction is defined on the basis of signal dynamics, generally at p% of the signal power variation range, i.e.,

$$P_{th} = P_{\min} + p \cdot (P_{\max} - P_{\min})/100 \cdot \quad (3)$$

The power dynamics in longer utterances is often variable and the above mentioned threshold has to be updated on frame basis. This is done again by exponential averaging:

$$P_{\max[\min]}(i) = q \cdot P_{\max[\min]}(i-1) + (1-q) \cdot \overline{P}(i), \quad (4)$$

where the speed of the averaging controlled by the parameter q differs according to the given context. Typically, the increase of $P_{\max}$ and the decrease of $P_{\min}$ must be quite fast (q ~ 0.9) while the decrease of $P_{\max}$ as well as the increase of $P_{\min}$ should be slower (q ~ 0.999). Such adaptive thresholding is demonstrated in Fig.1. To avoid failures of the algorithm when $P_{\min}$ and $P_{\max}$ are too close, their minimal distance can be defined. When reached, further decrease of $P_{\max}$ or the increase of $P_{\min}$ is not allowed.

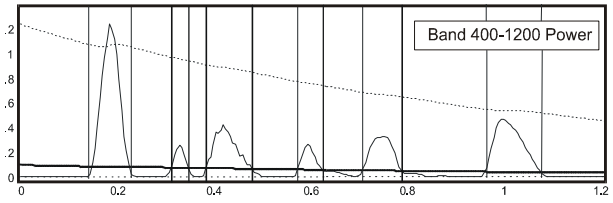

Figure 1: *Illustrative graph of energy in 2nd frequency band with the adaptive threshold updating*

3. Results

3.1. Study I

Experimenting with various settings showed that the optimal performance can be achieved with the band of 400-1200 Hz. The adaptive threshold for this band was set at 8 %. With this setting, values of the Czech speaker clearly differed from those of the English speaker. Neither the overall energy across the spectrum nor various other frequency bands provided useful results. Table 1 presents two simple confusion matrices from linear discriminant analyses performed on manually and automatically extracted parameters. The manual data yielded success rate of 82.4 %, while the automatically extracted data led to success of 80.6 % correctly recognized breath-groups.

Table 1. *Confusion matrices from discrim. analyses of manually (M) & automatically (Au) extracted data.*

n = 108	English-M	Czech-M
English	42	11
Czech	8	47
n = 108	English-Au	Czech-Au
English	42	11
Czech	10	45

The incorrectly recognized rhythm in about 20 % of the items was caused by local variation in individual breath-groups. Given that we are dealing with 'global rhythm measures' we decided to congregate the breath-groups by five. This created larger stretches of speech and neutralized local deviations from the overall rhythmic structure. (5 breath-groups typically stood for 1 paragraph of news.) The original 108 breath-groups were put into 20 paragraph-size-groups. This step made the English and Czech material separable with 100% accuracy and the overlap between the two languages vanished.

The results of Study I also showed that automatically extracted high-energy (HE) intervals were not the same as manually labeled vocalic (V) intervals and low-energy (LE) intervals were not equal to the consonantal (C) intervals. This is what we expected since the primary motivation of our work was to deal with the ambiguous status of sonorous consonants. Some of the [r], [l], [j], [w], [m], [n], [ɲ] or [ŋ] phones were treated according to their actual energy profile in the given context. Thus they may have contributed to the rhythm pattern regardless their phonological status.

One of the consequences of the different treatment of sonorants was a shift in the 'typological portrait' of the compared languages. As shown in Fig. 2, the manual labeling portrays Czech as a language with smaller proportion of V intervals in speech (%V) while automatic extraction of the energy profile shows Czech to have greater ratio of high energy intervals in speech (%HE). Otherwise, the consonantal (ΔC) and low-energy interval variation (ΔLE) behave analogically (and so do the pairwise variability measures PVI-V with PVI-HE and PVI-C with PVI-LE, which were not plotted in this paper for the lack of space).

3.2. Study II

Study II was carried out to check the consistency of the emergent pattern. The results are relatively straightforward. Four Czech and four English newsreaders produced 464 breath-groups which were congregated into 93 groupings by five consecutive breath-groups. Fig. 3 shows their properties along all four extracted parameters based on durations of high-energy (HE)/low-energy (LE) intervals.

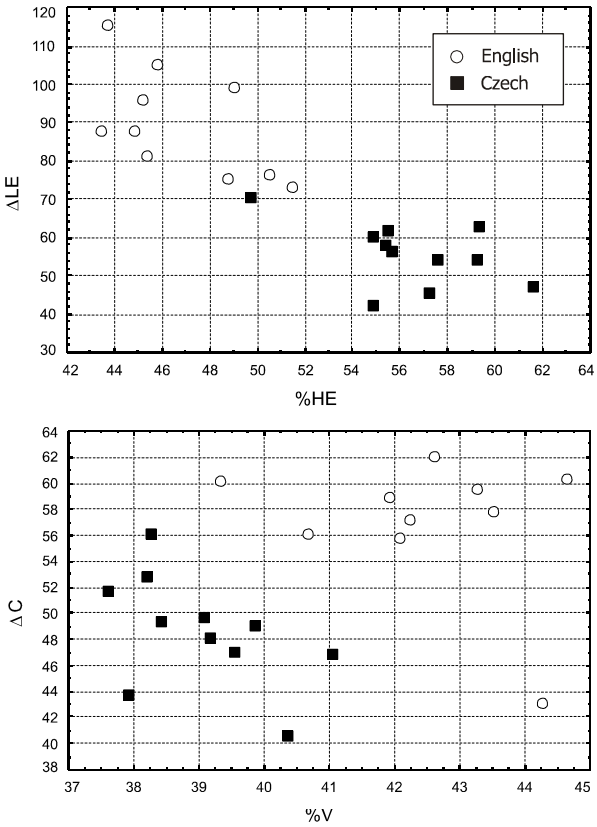

Figure 2: *Differences between Czech and English in High-Energy/Low-Energy metrics (upper plot) and the Vowel-Consonant metrics (lower plot).*

It is obvious that Czech HE intervals tend to occupy greater portion in speech (%HE) and they are less variable (PVI-HE). The English LE intervals are more variable in terms of standard deviation from the mean (ΔLE), but not in terms of pairwise variability index (PVI-LE).

The discriminant analysis was carried out using three relevant variables: %HE, ΔLE and PVI-HE. The whole set of 93 items was divided into 47 training and 46 testing items by random choice. The training items provided the classification functions which were used on the testing set. They classified all the items correctly as English or Czech apart from two in each language group. This implies the success rate of 91.3%.

4. Discussion

The results achieved so far are quite promising. It seems that energy distribution in a restricted frequency band of the speech signal can indeed reflect global rhythmic attributes of the language which is being spoken. Nevertheless, the current results should be treated with caution. First, news reading represents a very specific speech style. It remains to be seen whether the results can be further generalized. Former research revealed that not only various styles, but even different tempos sometimes produced quite disparate rhythm measures.

Second, we agree with [19] in that rhythm as a blanket term actually involves a number of interrelated phenomena. The metrics we have worked with in the current study reflect certain structural properties of the Czech and English languages and are not necessarily in any simple relationship with the perceived rhythmical patterns.

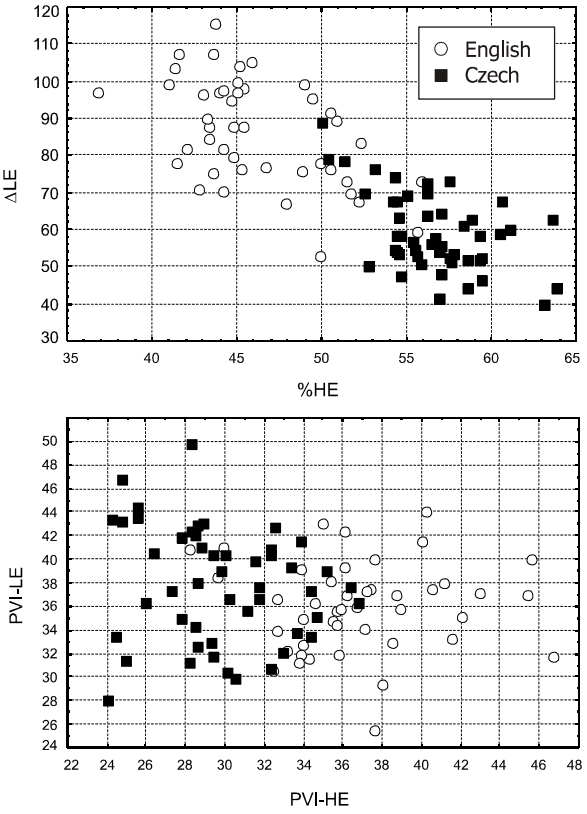

Figure 3: *Differences between Czech and English spoken texts in HE/LE metrics (n = 93; see text).*

Rhythm by definition is always primarily a perceptual matter. Therefore, it has to be emphasized together with [20: 334] that perceptual tests should be part of any endeavor to establish rhythm class membership or test the rhythm class hypothesis.

Third, it has to be stated explicitly that we were not looking for correlations with manual labeling. On the contrary, since we questioned its validity with regard to sonorant consonants, we expected that our approach will relocate the boundaries between the original consonantal and vocalic intervals. This indeed happened, yet our high-energy and low-energy intervals still consistently captured the difference between given Czech and English spoken texts. Whether they also reflect perceptual impressions to a greater degree than the manually labeled objects remains to be seen through further research.

5. Conclusions

Our study shows that global rhythm metrics based on durations of vocalic and consonantal intervals which are commonly obtained after careful manual labeling have their counterpart in measurements of speech signal energy in the 400-1200Hz band. The cutoff point which separates high-energy and low-energy intervals had to be established empirically. Our material required adaptive threshold set at 8 % of the local maximum.

The comparison of results based on traditional manual labeling with our automatically extracted parameters showed that the latter are capable of differentiating between Czech and English to the same extent as the former. However, the structural differences captured by automatically extracted data do not simply parallel the manual labeling. This fact should be at focal point of the future research.

Consistency of our findings was tested on the speech material from eight regular news bulletins. It seems that given speech style indeed exhibits different temporal patterns for Czech and English. The limits of further generalizability will be tested in the nearest future.

6. Acknowledgements

The first author's work was supported by the European grant MRTN-CT-2006-035561 "Sound to Sense". The second author gratefully acknowledges the support of the grant GACR 102/08/0707 "Speech Recognition under Real Word Conditions" and MSM 6840770014 "Perspective Informative and Communications Technicalities Research".

7. References

[1] Buxton, H., "Temporal predictability in the perception of English speech", in A. Cutler and D. R. Ladd [Eds], Prosody: Models and Measurements, 111-121, Springer-Verlag, 1982.

[2] Grossberg, S., "Resonant neural dynamics of speech perception", Journal of Phonetics 31: 423-445, 2003.

[3] Port, R., "Meter and speech", J. of Phonetics 31, 599-611, 2003.

[4] Cummins, F., "Rhythm as entertainment: the case of synchronous speech", Journal of Phonetics 37, 16–28, 2009.

[5] Marcus, S.M., "Acoustic determinants of perceptual centre location", Perception & Psychophysics 30, 247-256, 1981.

[6] Pompino-Marschall, B., "On the psychoacoustic nature of the P-centre phenomenon", Journal of Phonetics 17, 175-192, 1989.

[7] Howell, P., "Prediction of P-centre location from the distribution of energy in the amplitude envelope", Perception & Psychophysics 43, 90-93, 1988.

[8] Fowler, C.A., Whalen, D.H. & Cooper, A.M., "Perceived timing is produced timing: A reply to Howell", Perception & Psychophysics 43, 94-98, 1988.

[9] Ramus, F., Nespor, M. and Mehler, J., "Correlates of linguistic rhythm in the speech signal", Cognition 73, 265-292, 1999.

[10] Grabe, E. and Low, E.L., "Durational variability in speech and the rhythm class", in C. Gussenhoven & N. Warner [Eds.]: Papers in Lab. Phonology 7, 515-546, Mouton de Gruyter, 2002.

[11] Barry, W. J. et al., "Do rhythm measures tell us anything about language type?", Proc. 15th ICPhS, 2693-6. Barcelona, 2003.

[12] Gibbon, D. and Gut, U., "Measuring speech rhythm", Eurospeech Proceedings, 91-94, Aalborg: ISCA, 2001.

[13] Wagner, P. and Dellwo, V. "Introducing YARD and re-introducing isochrony to rhythm research", Speech Prosody Proc., Nara: SProSIG, 2004.

[14] Dankovičová, D. and Dellwo, V. "Czech speech rhythm and rhythm class hypothesis". Proc. 16th ICPhS, 1241-1244. Saarbrücken: IPA & UDS, 2007.

[15] White, L., Mattys, S., Series, L. and Gage, S. "Rhythmic metrics predict rhythmic discrimination" Proc. 16th ICPhS, Vol. II, 1009-1012. Saarbrücken: IPA & UDS, 2007.

[16] Asu, E.L. and Nolan, F., "Estonian and English rhythm: a two-dimensional quantification based on syllables and feet", Speech Prosody Proc., Dresden: TUD, 2006.

[17] Pollak, P., Rajnoha, J., "Long recording segmentation based on simple power voice activity detection with adaptive threshold and post-processing", SPECOM 2009, St. Petersburg, 2009.

[18] Xie, Z., Niyogi, P., "Robust acoustic-based syllable detection", Interspeech 2006 - ICSLP, 1571-1574, Pittsburgh, 2006.

[19] Barry, W. J., "Rhythm as an L2 problem: How prosodic is it?", in J. Trouvain and U. Gut [Eds], Non-native prosody, 97-120, Mouton de Gruyter, 2007.

[20] Kim, J., Davis, Ch. and Cutler, A., "Perceptual tests of rhythmic similarity: II. Syllable rhythm", Language and Speech 51/4, 343-359, 2008.