# Temporal Downtrends in Czech Read Speech

*Jan Volín*, *Radek Skarnitzl*

Institute of Phonetics, Charles University in Prague - Faculty of Arts, Czech Republic

{jan.volin, radek.skarnitzl}@ff.cuni.cz

## Abstract

A possible existence of a regular temporal trend superimposed over the durational pattern of individual segments is explored in read continuous speech in the western Slavonic language of Czech. A short text read by 75 speakers was used to ascertain whether the contextually conditioned temporal variation would allow any phrasal tendencies to manifest. The data were normalized against the speakers' characteristics and against the intrinsic duration of individual phones. The results indicate that while the linear trendlines are regularly declining, the most reliable partial trend is phrase-final deceleration. Three more general non-linear trends are identified.

**Index Terms**: temporal modelling, duration, downtrend, phrasal component

## 1. Introduction

Substantial evidence exists that quasi-universal phenomena concerning temporal patterning of speech are modified in every language in a distinct manner. There is reason to believe that such language-specific modifications are not random, but that they reflect constraints given by, e.g., the morphology (analytical vs. synthetic), phonological contrast (various levels of lexical density), or the phonotactics (complex vs. simple) of the language. Naturally, this conviction has yet to be detailed by thorough research of individual languages.

Temporal modelling of speech often builds on the simple physical fact that duration is in a direct inverse relationship with rate. Thus, a shorter duration of units in a chain results in a higher speaking rate and vice versa. This approach is quite useful especially from the viewpoint of concatenative synthesis. To specify durations of individual segments (i.e., phones) means to account for speech tempo as such, since one is determined by the other. In his classical study, Klatt maintained that segmental duration in English provided cues for the distinctions between (1) inherently long and short vowels, (2) fortis and lenis consonants, (3) phrase-final and non-final syllables, (4) fortis/lenis postvocalic consonants through the duration of the preceding vowel, (5) accented and unaccented syllables, and (6) the presence or absence of emphasis [1]. While (1), (2) and (4) are related to segments as such, (3), (5) and (6) concern the prosodic layer of speech.

Clearly, there is no argument that durational information is also perceptually important with regard to the structure of the utterance. A shorter duration of a speech segment can help to distinguish it from an intrinsically longer one while, simultaneously, providing information about the local articulation rate which might, in turn, reflect the prosodic structure of the utterance. Klatt [1] devised a series of recursive formulae that predicted durations of English vowels and consonants through multiplication of a base value by coefficients related to various segmental and suprasegmental attributes.

For the sake of linguistic theory, however, it seems quite reasonable to ask whether segmental and suprasegmental temporal features are controlled one through the other or whether they operate on autonomous levels. More specifically, one might want to explore the possibility that something like a phrasal component (similar to the one proposed by H. Fujisaki for intonation description) or to a prosodic gesture (the time-warping π-gesture as in [2]) manifests itself on the level of a prosodic unit when other, more local durational influences are factored out. If such a prosodic component or gesture exists, then it might be useful to know its parameters for a given language, speaking style, or perhaps even a given speaker. Knowing the parameters of the hypothetical temporal course would enable us to build models with many practical applications.

Figure 1 (overleaf) displays some of the possible models of the phrasal temporal component in an unspecified domain. The figure does not exhibit the so-called zero model, which would basically amount to a straight line, suggesting that articulation rate is kept constant throughout the domain. Model a) actually also proposes such constancy but it adds the final deceleration also known as final lengthening. Model b) suggests a gradual decline in rate throughout the whole unit. The decline is steeper in the final portion of the unit. Such model would resemble a global intonation trend found in some languages. Model c) would predict gradual acceleration when departing from the prosodic boundary and deceleration when approaching it. Although quite logical, such a neat, symmetrical outcome does not seem to be supported by empirical data. Asymmetry as seen in version d) would appear more realistic, since it reflects on the evidence mentioned in [2: 166]. The last model contains some of the features of the previous: final deceleration is more moderate than initial acceleration, the initial part of the unit contains gradual decline and there is also a portion of constant rate in the mid and near final region. This model merely implies the possibility of combinations of various simpler temporal contours.

The question of perceptual impact of such temporal modifications will not be considered at this point. Suffice it to say that naturalness of speech pertains to many phenomena that are not directly accessible to conscious intellectual observation. Even such a conspicuous feature as final deceleration is most of the time registered on the subconscious level. Finer temporal modifications must be ascertained by more sophisticated methods. An attempt to do exactly that is presented in this study.

## 2. Method

### 2.1. Material

The material was based on a short meaningful text describing an interaction of a schoolboy with his grandmother. It consisted of 9 sentences comprising 51 words, 91 syllables, and 220 phones if pronounced canonically. Modal prosodic segmentation induced by the semantics and syntax of the text amounted to 14 intonation phrases.

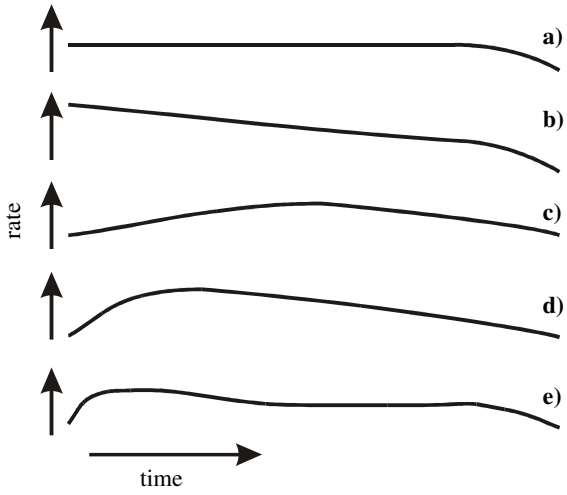


Figure 1: *A few models of the phrasal temporal component with time on the abscissa and rate on the ordinate.*

Seventy-five Czech university students (48 women and 27 men) were asked to read the story. They were instructed to read in a relaxed manner and as fluently and naturally as possible, and were given adequate time for preparation. After they had familiarized themselves with the text, they were recorded individually in a soundproof booth. They were prevented from listening to each other's production to avoid unintentional imitations. The recordings were made under identical conditions with an electret microphone IMG ECM 2000 and a soundcard SB Audigy 2 ZS.

The recordings were manually labelled by two experienced labellers [3]. The whole material consisted of about 3,800 words and 16,500 phones. Our experiment, however, builds on arithmetic means of durations of speech segments found in various positions. Therefore, we excluded phones that did not occur in seven or more different positions in the text. Also, post-pausal voiceless plosives were excluded as in [4]. Thus, the following analyses are based on 12,825 manually labelled phones.

### 2.2. Procedure

Raw durations of individual phones in milliseconds were converted into *articulation frequency* (*AF*) by calculating how many times a segment of the given duration (*dur*) would occur in one second:

$$AF = \frac{1}{dur} \tag{1}$$

We chose to call this variable articulation frequency (a segmental parameter) rather than articulation rate (*AR* - a suprasegmental phenomenon) since the latter is usually used in different contexts for different purposes.

In addition, the raw values in milliseconds were also normalized by conversion into z-scores and transformed to resemble the millisecond values:

$$AF_z = \frac{1}{30 \cdot \left( \frac{dur - \bar{x}}{s} \right) + 70} \tag{2}$$

$AF_z$ stands for z-transformed articulation frequency, *dur* is the duration of a segment again, $\bar{x}$ is the mean segmental duration in the given speaker's production and *s* is the standard deviation of this mean. Obviously, multiplication of the z-score by a constant and addition of a constant does not change the distribution of the values.

Another type of transformation which we tested was the one by syllabic articulation rate ratio. The number of syllables produced by the given speaker was divided by the duration of the text without pauses. This resulted in the average articulation rate in syllables per second ($AR_S$). We calculated the mean across the $AR_S$ of all of our 75 speakers. The ratio of the individual speaker's $AR_S$ to the overall mean $AR_S$ was dubbed *index AR* ($Id_{AR}$). The speaker with individual $AR_S$ equal to the overall mean $AR_S$ had $Id_{AR} = 1$. The index $Id_{AR}$ was then used to multiply the durations in milliseconds of the individual segments:

$$AF_S = \frac{1}{Id_{AR} \cdot dur} \tag{3}$$

The raw data in milliseconds and two sets of normalized data were used to express the relationship between the average *AF* and the actual *AF* of a specific segment in two ways. The first was the real difference in the mathematical sense (*Actual AF* minus *Mean AF*). The second was the percentual ratio of the actual *AF* to the mean *AF* (100**Actual AF*/*Mean AF*).

One of the preliminary questions asked in this study was what effects the six procedures used (3 sets of data × 2 types of difference representation) will have on the results, e.g., which of them will reduce the variation caused by individual speaking rates most successfully.

## 3. Results

First of all, we compared the coefficients of determination from linear correlation analyses to find out which of the six procedures provided the greatest portion of explained variance. The results were actually very much the same for all six of them. Some slight advantage emerged in using z-score transformed data ($AF_z$) together with the percentual expression of the relation between the actual and the mean *AF*. This variable will be henceforth called *relative AF*. Results from this mode of analysis are displayed in Table 1.

| *IP* | *n SUs* | $r^2$ | *trend* |
|---|---|---|---|
| *1* | 3 | 0.31 | falling |
| *2* | 1 | 0.95 | falling |
| *3* | 3 | 0.43 | falling |
| *4* | 2 | 0.62 | falling |
| *5* | 2 | 0.40 | falling |
| *6* | 3 | 0.28 | falling |
| *7* | 4 | 0.20 | falling |
| *8* | 2 | 0.39 | falling |
| *9* | 2 | 0.01 | level |
| *10* | 2 | 0.67 | falling |
| *11* | 3 | 0.18 | falling |
| *12* | 3 | 0.21 | falling |
| *13* | 1 | 0.92 | falling |
| *14* | 3 | 0.12 | falling |

Table 1. *Numbers of stress-units (n SUs) in individual intonation phrases (IP), coefficients of determination ($r^2$), and linear trends (trend) from analysis of z-score standardized data.*

The linear regression analyses showed that with one exception (the ninth intonation phrase – IP 9), all the IP trendlines were declining. It can be also observed in the table that there is a

relationship between the number of stress-units (SUs) in the phrase (IP) and the variance explained by the linear trend ($r^2$). The variance in the shortest IPs (IP 2 and 13) is almost completely captured by a straight line, while longer IPs have progressively less variance explained by the linear fit. The exception is again IP 9. Close inspection of this phrase revealed that it was actually subject to unusually great variation of prosodic phrasing by individual speakers. Also, 23 of the 75 speakers made errors or alterations when they read this part of the text due to the complex syntax and one four-syllable low-frequency word there. When we selected only the speakers who read IP 9 fluently and with the prosodic break as expected, the trend became also falling. In our further search for the phrasal component of the temporal variation, we will use the corrected data for IP 9.

Having established the linear trend in the course of relative *AF*, we looked at the initial and final values in each of the phrases to see whether there was any regularity. Table 2 shows the values rounded to the nearest multiple of 5. As for the final segments, there was a considerable concentration of values in the region of 45 to 70 %. In other words, the final segments in our phrases tended to be slowed down to about 60 % of their mean AF.

| *IP* | *initial (%)* | *final (%)* |
|---|---|---|
| *1* | 120 | 60 |
| *2* | 120 | 70 |
| *3* | 115 | 45 |
| *4* | 180 | 70 |
| *5* | 125 | 60 |
| *6* | 135 | 50 |
| *7* | 125 | 65 |
| *8* | 125 | 55 |
| *9* | 75 | 65 |
| *10* | 110 | 50 |
| *11* | 180 | 70 |
| *12* | 135 | 65 |
| *13* | 135 | 60 |
| *14* | 90 | 60 |

Table 2. *Relative articulation frequencies (AFs) of the initial and final segments in each phrase.*

The range of the initial values is markedly larger, yet ten of the values lie in the region of 110 to 135 %. The two initial values of 180 % in phrases 4 and 11 belong to vowels in monosyllabic structural words (a pronoun and a conjunction respectively), while the two slow values in phrases 9 and 14 belong to consonants in content words after rather shallow prosodic boundaries. Phrase 14 is, furthermore, the concluding sentence of the story and all the *AF* values in it were relatively low.

Knowing the initial and final values and the linear trend of the phrase allows us to build only a very simplified model of the temporal change within an IP. This was already suggested by the difference between the linear fit of shorter and longer IPs (see above Table 1). Therefore, our next step was to fit the data with a non-linear 'trend-curve' to achieve greater precision in mapping the temporal course throughout the phrase. We opted for polynomial functions and the least-sum-of-squares method. The one-stress-unit IPs 2 and 13 were not processed as they exhibited almost perfect linear fit. For the other twelve phrases, the non-linear regressions led naturally to improved models, as indicated by the difference between the third and fourth columns in Table 3.

| *IP* | *n SUs* | *lin. $r^2$* | *non-lin.$r^2$* | *trend* |
|---|---|---|---|---|
| *1* | 3 | 0.31 | 0.54 | B |
| *3* | 3 | 0.43 | 0.86 | B |
| *4* | 2 | 0.62 | 0.72 | C |
| *5* | 2 | 0.40 | 0.60 | C |
| *6* | 3 | 0.28 | 0.53 | B |
| *7* | 4 | 0.20 | 0.42 | A |
| *8* | 2 | 0.39 | 0.67 | A |
| *9* | 2 | 0.06 | 0.54 | B |
| *10* | 2 | 0.67 | 0.74 | A |
| *11* | 3 | 0.18 | 0.29 | A |
| *12* | 3 | 0.21 | 0.38 | A |
| *14* | 3 | 0.12 | 0.76 | B |

Table 3. *Numbers of stress-units (n SUs) in individual IPs, coefficients of determination for linear (lin. $r^2$) and non-linear (non-lin. $r^2$) models, and types of the overall trend (trend).*

Three overall trends which were found in the data are portrayed in Fig. 2. Five of the cases displayed temporal pattern A, which means short deceleration at the beginning of the phrase, moderate deceleration or level tempo in the middle, and quite substantial deceleration at the end. Type B, which was also found in five IPs, differed only in that the beginning of the phrase did not display the short gradual decrease in tempo, but the opposite. Type C occurred twice. It had one or two fast segments at the beginning followed by some fluctuation and a final decline in tempo. The final acceleration glitch is actually caused by consonantal codas which were slightly faster in comparison with the preceding vowels. Types A and B, on the other hand, ended with open syllables (except for IP 10, which had a nasal coda).

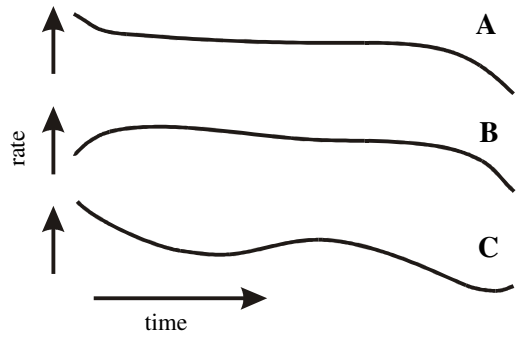


Figure 2: *Three non-linear trends of the phrasal temporal component found in the data (time on the abscissa, rate on the ordinate).*

It is also worth noticing that while types A and B were not bound to IPs of certain length, type C occurred exclusively on IPs consisting of 2 SUs. It would be tempting to look at type C as two slopes corresponding to two SUs. This, however, would be incorrect. The pre-final peak in tempo does not correspond to any particular part of a stress-unit. In IP 4 it occurs on the second syllable of the second four-syllable SU, while in IP 5 it is located just before the last monosyllabic SU. Thus, local influences overrule the prosodic structure of the whole.

The same might be true about the difference between types A and B. The former starts with a faster *AF*, while the latter with a slower one. There are two possible causes of this state. Type-A phrases start with monosyllabic structural words (with one disyllabic exception) and the fast elements in them are

vowels. Type-B phrases start with content words and the downward pulling elements are consonants. We can see that vowels behave differently from consonants with respect to the temporal course in the higher units. Although Czech is a language without vocalic reduction in its standard form, there is clear durational reduction of vowels in some structural words.

Since we were looking for the phrasal component, we wanted to further process the data with regard to those influences that are generally independent of IP structure. We carried out multiple regression analysis with the relative *AF* as the dependent variable. Independent variables were the following: the position of a segment in the SU, the position of the SU in the IP, the size of the SU, the semantic status of the word, and text frequency of the word. The position of the SU in IP was clearly the most important predictor in the model ($\beta$ = -0.425; $p < 0.001$). However, it was also the one which we did not want to factor out if we were to preserve the phrasal temporal component.

As for the other independent variables, the multiple regression analysis confirmed what we noticed in our previous observations: the significant effects were found only for vowels. Relevant predictors were the semantic status of the word ($\beta$ = 0.304; $p < 0.01$) and the position of the segment in the SU ($\beta$ = -0.349; $p < 0.001$). The size of the SU was *not* significant, which confirmed the findings of [5] for similar speech material. Word frequency was also found insignificant in our material. As it was correlated with the semantic status of the word, however, it became significant in slimmer models where the semantic status was excluded. Its significance was still lower than that of the semantic status of the word. The size of the SU remained insignificant in all the combinations of variables tested.

Since the consonants appeared quite inert with respect to all the descriptors in the analysis, further processing affected only the *AF* of the vowels. We first used the regression coefficient of the semantic status of words to transform the vocalic *AF* values in order to see whether the phrasal temporal structure can be made clearer. The transformation led to a considerable reduction of the total sum of squared residuals: from 42,790 to 35,767 (ΣSS reflects the scatter of data points in space). The results, however, did not bring any significant change in the general shape of the curves or in the fit of the curves to the data points as expressed by $R^2$. The same outcome emerged from modifications in which the regression coefficient of the segment position in the SU was employed. Thus, although the overall temporal trend in IPs is not random, the goodness of fit seems to be difficult to influence by general rules. The local positional and contextual effects are obviously intertwined in a non-trivial manner.

## 4. Discussion and conclusion

The linear trends through the individual data points in all the 14 intonation phrases we worked with were falling (i.e., had negative gradients). There was also a considerable stability in the values of the initial and especially the final segments in the phrases. The non-linear regressions showed that the temporal downtrend in the phrases was, to a large extent, driven by the final deceleration and in more than half of the cases by fast segments at the beginning of a phrase.

Multiple regression analysis revealed that the fast pronunciation of structural words (as opposed to content words) was a matter of durational reduction of vowels. Consonants were not affected by the semantic status of the word in which they occurred. However, an attempt to use regression coefficients to improve the fit of the non-linear model did not bring about substantial changes. The overall scatter of the data was reduced, but we did not find any general rules that would improve the fit across the board. For example, taking care of vowel reduction in structural words improved the situation for 12 of them, while the remaining 7 responded in an undesirable way. The trouble is that these two groups of structural words did not differ in any meaningful sense – they both consisted of pronouns, prepositions and auxiliary verbs. Conjunctions were found only in the 'good' group. Similarly, vowels in polysyllabic stress units tend to be longer in SU-final and shorter in SU-initial positions, but there were many exceptions from this pattern, which, most probably, stem from interactions of individual neighbouring segments. Therefore, we are unable to compensate for certain inconsistencies at this stage.

The strength of local positional and contextual influences and the number of unique rules governing mutual impact of individual segments on one another is not surprising to scientists who build durational sum-of-product models and who could provide many detailed examples. Even before such models, more than 30 years ago, Umeda [6] showed that each vowel of the American English system had its own pattern of elongation or shortening under different conditions. It follows that simple coefficients which should apply to all vowels in a multiplicative model have to be modified according to the individual properties of the segment under consideration.

Nevertheless, as we had another short text read by the same speakers, we wanted to know if the relative *AF* of segments in the same position and context could be predicted. To find comparable positions and contexts in short (one-paragraph) texts is difficult, but we managed to identify a vowel [a], which in both texts lay in a three-syllable penultimate stress-unit followed by a two-syllable phrase-final SU. The vowel was preceded by [l] and followed by initial [p] of the next word. According to our prediction, the vowel should have relative *AF* of 110.5 % of its mean *AF*. In reality, the observed value was 108 %. The mismatch of only 2.5 % could be considered a very good result.

In conclusion, we can say that the Czech language possesses a relatively reliable phrasal temporal component. However, further research is essential if we want to learn more about its structure and stability.

## 5. Acknowledgements

This research was supported by the grants VZ MSM 0021620825 and GACR 405/05/0436.

## 6. References

[1] Klatt, D.H., "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence", J. Acoust. Soc. Amer., Vol. 59, 1976, 1208-1221.

[2] Byrd, D., Saltzman, E., "The elastic phrase: modelling the dynamics of boundary-adjacent lengthening", Journal of Phonetics, Vol. 31, 2003, 149-180.

[3] Volín, J., Skarnitzl, R., Pollák, P., "Confronting HMM-based Phone Labelling with Human Evaluation of Speech Production". Interspeech 2005 Proc., 1541-1544, 2005.

[4] Moebius, B., van Santen, J. P. H., "Modelling segmental duration in German text-to-speech synthesis", ICSLP96 Proc., 2395-2398, 1996.

[5] Dankovičová, J., The linguistic basis of articulation rate variation, Hector, Frankurt am Main, 2001.

[6] Umeda, N., "Vowel duration in American English", J. Acoust. Soc. Amer., Vol. 58, 1975, 434-445.