Phrase Time Structure Modeling for Speech Synthesis Purposes

Grażyna Demenko

Institute of Linguistics
A. Mickiewicz University, Poznań, Poland
lin@amu.edu.pl

ABSTRACT

The paper presents a model of the intonational – rhythmic phrase structure in the Polish language as well as the premises for speech sound duration analysis to be used in text-to speech synthesis. The model of the prosodic phrase has been tested on the acoustic and perceptual analysis of the 50 syntactically and semantically diversified utterances produced by 40 native Polish speakers. The results showed that tempo changes within a phrase and the locus for a change in tempo is the focus. The statistical analysis showed the importance of the separate rhythm modeling in the prenuclear and nuclear part of the phrase. The results showed the possibility of timing modeling with the 76-87% correctness depending on the complexity of the phrase. The phrase time structure modeling will be tested in the actually build synthesizer of Polish speech.

1. INTRODUCTION

The time structure of utterances has an important influence on the quality of synthesized speech. Modeling acoustic-phonetic segment durations presents a complex task and so far it has not been satisfactorily tackled for the purposes of practical applications. In the majority of studies on prosody modeling for speech synthesis purposes most attention is paid to intonation while duration and speech rhythm seem to be regarded as second-rate factors. The necessity of considering the factors on the suprasegmental level is emphasized in many papers on duration, however, there is a deficiency of comprehensive investigation into the relations between intonation and the rhythm structure of utterances. The reasons of the lack of adequate descriptions of utterance time structure are twofold – firstly, practical, related mainly to the speech signal segmentation, and secondly, theoretical, related to the difficulties in determining the sources of variability influencing the duration of specific acoustic-phonetic segments.

The duration of individual segments within an utterance is modified by the interactively functioning linguistic and extralinguistic sources of variability (Klatt [8], Campbell [1]). The durational structure of a sentence is influenced by the semantic, syntactic factors related to the position of a sentence /phrase/ word within text, rhythm and segmental effects. The influence of syntactic factors observed in experimental studies manifests itself in substantial lengthening (up to a few dozen per cent) of the final syllables, mainly of the last vowel or both the vowel and sonorant consonants or fricatives following the vowel. A similar result was obtained for Polish (Demenko [2]). The segmental durations in read texts were analyzed. Four factors realized on two, three or four levels were finally taken into account, namely: the inherent phonetic segment duration, the stress placement, the syllable position within a phrase, and the structure of the phrase-final syllable. The analysis of the above factors showed that the vicinity of phrase boundary exerts the greatest influence on sound duration. The vowel lengthening in the final and in the penultimate syllable of the phrase performs the function of informing listeners of the syntactic and semantic structure of the utterance.

Speech rhythm is one of the variously defined factors determining sound duration. In many languages speech rhythm is related to the phenomenon of isochrony consisting in keeping relatively constant length of rhythm units, regardless of the number of syllables (Lehiste [9], Sagisaka et al. [11]). In the isochronic languages (e.g. English, German, Dutch, Polish) speech sounds are shortened to a certain degree along with the lengthening of the rhythm unit. Jassem [7] presented an analysis of isochrony useful for practical applications. He drew up two types of rhythm units on the basis of a regression model: NRU - the narrow rhythm unit and ANA – the anacrusis. NRU depends on the number of syllables. The complete rhythm unit comprises both the narrow rhythm unit and the anacrusis. One of the more interesting solutions to the problem of duration modeling for the TTS purposes is the Prosynth model based on the phonological theory (Ogden et al. [10]), which is a syllable-based model. Simplified isochrony models based on vowel duration analyses are also developed e.g. the PVI - Pairwise Variability Index (e.g. Gibbon [6], Grabe [5]). So far, however, none of the existing utterance structure models has permitted of unconstrained speech rate and speech rhythm control.

2. INTONATIONAL-RHYTHMICAL PHRASE MODEL FOR POLISH.

Figure 1 presents an intonational-rhythmical phrase structure proposed for Polish. The duration of phonetic segments is repeatedly modified on different levels of the utterance production. The structure of the prosodic model is

ISBN 1-876346-48-5 © 2003 UAB

hierarchic. The intonational phrase (I) is assumed to be the largest unit. The intonational phrase is determined by the optional pre-nuclear intonation (PI) and the obligatory nuclear intonation (NI). The pre-nuclear as well as the nuclear intonation structure is determined by the accentual groups (PA_i, NA, NDA), which carry the secondary real accent or the primary real accent. The stress groups consist of rhythm units (PP_i, NP_i) determining the rhythm of the utterance. Each of the rhythm units comprises a sequence of vowels and consonants. The intonational phrase is conditioned by the discourse type and by the position within discourse, hence it determines the general intonational and rhythmical structure of the utterance. Modeling of various speech styles requires flexibility in defining prosodic features. The dynamics of the speech rate changes in different fragments of the utterance, as well as the tone level changes decide about grouping words into phrases. The phrase boundary may be interpreted variously depending on the speech rate (the slower the speech rate is, the more phrase boundaries arise).

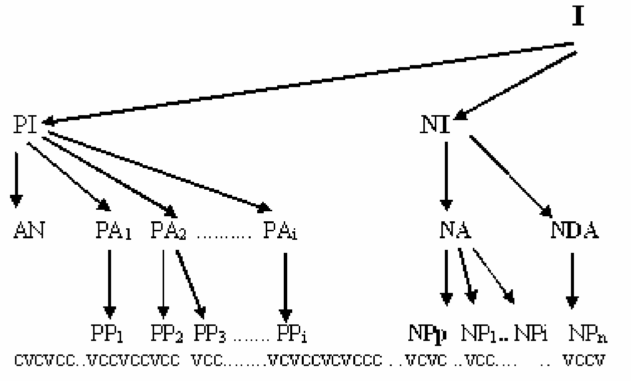
I
PI
NI
AN
PA1
PA2 ……….
PAi
NA
NDA
PP1
PP2
PP3 …….
PPi
NPp
NP1.. NPi
NPn
CVCVCC..VCCVCCVCC VCC…….VCVCCVCVCCC . .VCVC ..VCC….   .. VCCV

Fig.1: General structure of the intonational-rhythmical phrase model for Polish.

The pre-nuclear intonation may contain the anacrusis (a sequence of unstressed syllables – AN), and the pre-ictic intonational accents (PA_i). As a degenerate stress group, the anacrusis lacks the nucleus, it cannot carry any accent, and it can only occur in the phrase-initial position. At least two types of the pre-ictic accents are distinguished, as regards the pre-nuclear intonation: low and high. These are intonational accents and their phonetic description still has not been formulated precisely enough. The nuclear intonation is determined by the ictic accent – the primary real accent (NA) and the optional post-ictic durational accent (NDA). The primary real accent is the most important element of the intonational phrase. It determines various nuclear intonations depending on the style and the type of discourse. Nine nuclear intonation types have been distinguished for Polish (Demenko [3]). The ictic accent may be followed only by a non-pitch durational accent. The boundary between the pre-nuclear and the nuclear intonation is vague, and it may be localized within the intervocalic consonant group. The accentual groups' (PA_i, NA, NDA) boundaries are also vague. The number and distribution of accents depend on the general phrase structure. It was shown (Demenko [3]) that for Polish it is the vowel that is the most important to localize the accent. The fundamental frequency changes in the intervocalic consonants are not systematic. Feet are regarded as rhythm units the most frequently. Traditionally, the foot is defined as a sequence of syllables. Even though the notion of syllable is controversial, and it is particularly difficult to define its boundaries for Polish, in the light of the recent research for Polish (Demenko [3]) and for English (Santen [12]), it seems proper to determine the beginning of the vowel as the beginning of the rhythm unit. A question arises, however, where to assign the intervocalic consonants. Santen's results [12] indicate the necessity of separating the so-called rhyme (the vowel together with the following consonants) from the syllable.

In the phrase model for Polish presented here, the rhythm unit is understood as a fragment of signal beginning with a vowel (the potential stress carrier) containing a sequence of consonants and / or vowels and ending with the sound directly preceding the vowel being the potential carrier of the next stress (in Polish the word stress usually falls on the word's penultimate syllable). The rhythm units PP_1, PP_2, … PP_i are defined in the domain of the pre-nuclear accents. The NP_p is the first and the main unit in the nuclear accent, the NP_1 ... NP_i units following the NP_p are optional and they shape the nuclear intonation structure. The NP_n unit is optional and it determines the possible NDA accent.

Segmental duration is related to physiological conditions affecting the manner of signal production and the vocal organ functioning. In the intonational-rhytmical phrase model for Polish the following time structure modeling levels have been adopted:

1. Suprasegmental level - the intonational phrase level: the type of discourse determining speech rate, the phrase position within discourse, the complexity of the pre-ictic and ictic structure.
2. Suprasegmental level - the accentual group level: the group position (ictical, pre-ictical, post-ictical), the number of groups, the group structure (the number of feet).
3. Suprasegmental level - the rhythm unit level: the rhythm unit position (anacrusis, in the pre-ictic, ictic or post-ictic accent), the lenght of the unit.
4. Segmental level - the acoustic-phonetic segment level: sound type (vowel/consonant), the type of sound segment connection, the influence of the consonantal context, after a vowel and after a consonant.

3. EXPERIMENTAL RESEARCH

The linguistic material was comprised of more than 50 of syntactically and semantically differentiated phrases: questions, orders, and statements; e.g. *podaj adres* (Eng.: *give me the address*), *gdzie mam przesiadkę* (Eng.: *where do I change*), *o której przyjeżdża pociąg z Łomży do Wałbrzycha* (Eng.: *at what time does the train from Łomża to Wałbrzych arrive*). Forty speakers participated in the experiment; they were not recommended to use any particular way of speaking. The aim of the experiment was to practically assess the rhythm unit definition adopted in point 2 above. All utterances were subjected to the perceptual analysis, and the segmental durations were calculated on the assumed modeling levels. The utterances produced by most speakers with a complex intonation structure were selected for the detailed analysis (e.g. *podaj mi godzinę odjazdu pociągu z Kielc do Ostrowa*; Eng.: *give me the departure time of the train from Kielce to Ostrów*). The utterances produced with a pause (a break in the signal longer than 40 ms, except from the stop segments in stop consonants) were excluded from the analysis. The total duration of the phrases accepted for the analysis ranged from 2140 ms to 3664 ms. Table 1. presents the percentage duration of each of the obtained rhythm units calculated with the formula:

$$T_{Fi} = D_{Fi}/TD *100$$

where:

T_{Fi} - normalized rhythm unit duration (per cent), D_{Fi} - rhythm unit duration (ms), TD - total utterance duration (ms).

The analyzed units contained different number of phonemes even within the same utterance, e.g. in the utterance 1: (1) *odajmigodz'* (9 phonemes), (2) *ineodj* (6 phonemes), (3) *azdupotc'* (7 phonemes), (4) *oŋgusc* (6 phonemes), (5) *eldzdoostr* (9 phonemes), (6) *ova* (3 phonemes). The lengths of the accentual units appeared to be proportional to the number of their component phonemes. The greatest length diversity was observed in unit 5: *eldzdoostr*. For the majority of utterances the nuclear accent fell on the fragment *ova* in the word *ostrova* - the last word of the phrase (e.g. Fig. 2c, 2d). In most cases it was the ML nuclear accent (Fig. 2c), but some speakers produced it as HL (Fig. 2d). In three cases the nuclear accent was located on the word *Kielc* (e.g. Fig. 2a, 2b, 2e). A strong tendency for expansion of the pre-focal region was evident, particularly if the focal word is located near the beginning of the phrase. Post-focal regions are compressed (excluding final syllables). This findings, however, requires further support from extensive statistical analyses. Tempo modulations have been observed in Swedish as noted by Fant, Kruckenberg [4].

A correlation analysis has been carried out for all the utterances in order to estimate the recurrence degree of the time structure determined by the length of the accentual units under examination. Each of the utterances was defined by a six-dimensional vector of percentage lengths of the rhythm units distinguished in the utterance. The correlation coefficient calculated in a 40*40 size matrix ranged from 68 to 98 %. This indicates a relatively constant distribution of the individual units' duration within the utterance.

Figure 3 presents the acoustical analyses of the utterance *podaj mi godzinę odjazdu pociągu z Kielc do Ostrowa* produced by two speakers for whom the rhythm units' lengths correlated the most strongly with one another (in this case, the correlation coefficient amounted to 98 %).

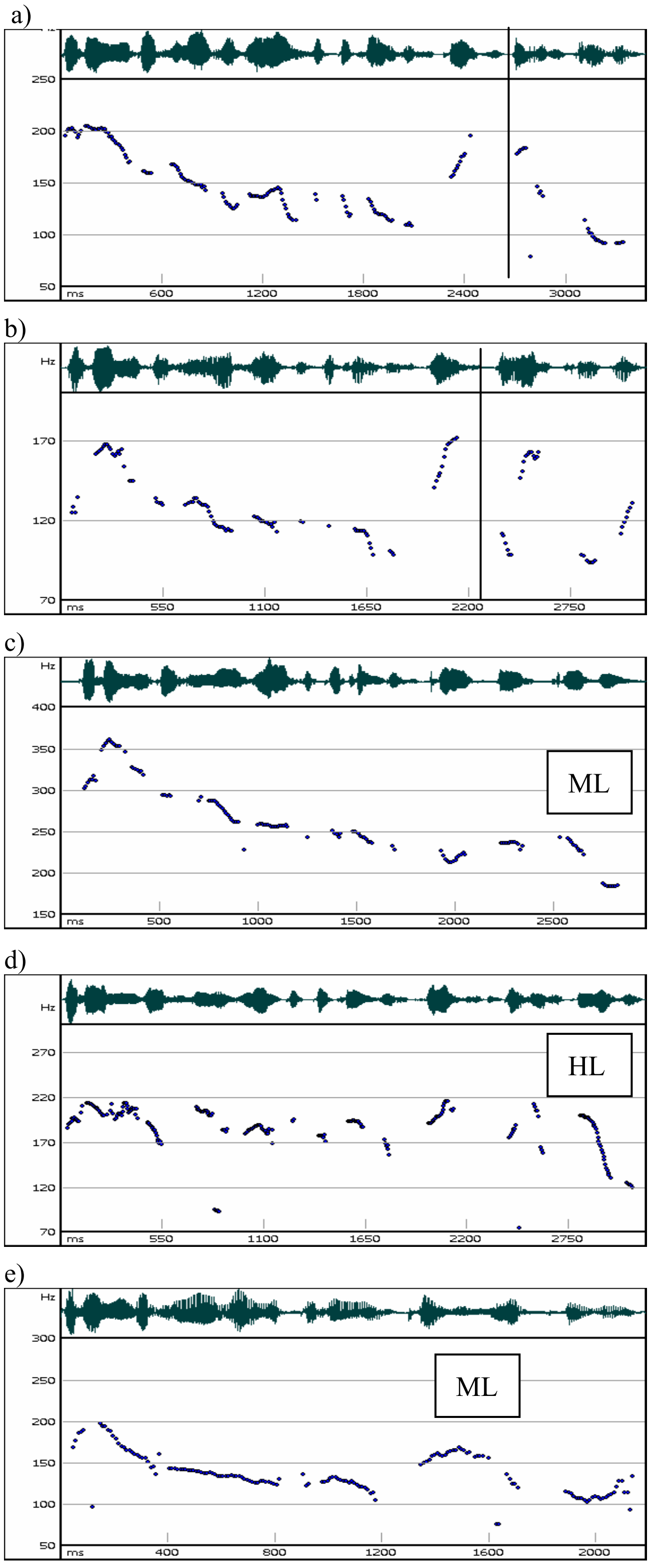

Fig. 2: Oscillograms and pitch contours for the utterance: *podaj mi godzinę odjazdu pociągu z Kielc do Ostrowa* a) phrase boundary after the word *Kielc*. The boundary has been marked with the coursor; b) first phrase boundary after *Kielc.* A different phrase structure; c) nuclear accent (ML) on the word *Ostrowa*; d) nuclear accent (HL) on the word *Ostrow*a; e) nuclear accent on the word *Kielc* (ML).

ISBN 1-876346-48-5 © 2003 UAB

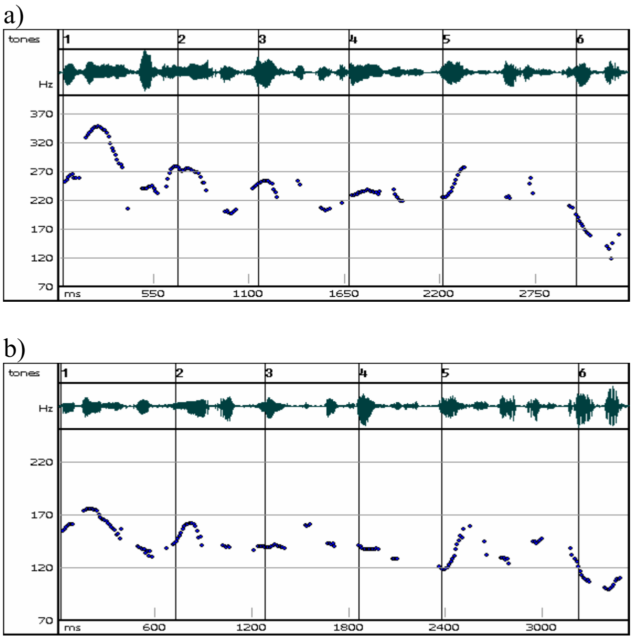

Fig. 3: Oscillograms and F_0 contours for the utterances produced by two speakers (a and b), for whom the rhythm units' lengths correlated the most strongly with one another. Coursors mark the boundaries of the 6 analyzed feet: 1- *odajmigodz'*, 2 – *inedj*, 3- *azdupotc'*, 4 – *oŋgusc*, 5 – *eldzdoostr*, 6 – *ova.*

	Lengths of the rhythm units (%)						Total ms
SR	19,6	15,2	17,4	14,2	22,5	10,5	3077
MN	16,8	11,3	14,7	12,1	18,9	7,7	2120
MX	23,4	18,0	21,6	17,3	27,4	13,2	3664
SD	1,4	1,6	1,4	1,2	2,2	1,3	325

Table 1: Mean lengths of the 6 rhythm units in the 5 utterances for 40 speakers: The four rows of the table contain basic statistics: the mean (SR), minimum (MN), maximum (MX) value and the standard deviation (SD).

4. CONCLUSIONS

1. The rhythm unit definition adopted in the study allows of systematic investigation of utterance time structure. A high recurrence of the time structures was observed for the analyzed utterance, regardless of individual differences in ways of speaking.
2. Time structure ought to be examined in association with intonation structure analysis.
3. The results show that there is a deceleration-acceleration pattern evident in focused phrase. The pre-focal regions in the analyzed sentences have been expanded and postfocal regions have been compressed.

The intonational-rhythmical phrase model proposed in this study requires detailed statistical analyses that would allow defining practical rules governing duration of specific acoustic-phonetic segments for speech synthesis purposes. The results will be applied in the currently conducted TTS project for Polish speech.

REFERENCES

[1] N. Campbell Synthesing Spontaneous Speech in *Computing Prosody* (Sagisaka Y., Campbell N., Higuchi N. ed.), Springer-Verlag New York, Inc., 165 – 185, 1997.

[2] G. Demenko *Acoustic classification and automatic recognition of accent and phrase boundary in speech signal,* Archives of Acoustics, vol.23, 2, 159 – 178, 1999.

[3] G. Demenko *Analysis of Polish Suprasegmentals for Speech Technology*, wyd. UAM, Poznań, 1999.

[4] G. Fant, G. A. Kruckenberg, L. Nord *Temporal organization and rhythm in Swedish,* Proceedings of the XII ICPhS, Aix-en-Provence, 242-245, 1991.

[5] E. Grabe, B.Post, I. Watson *The Acquisition of Rhytmic Patterns in English and French,* Proceedings of the 13th ICPhS,1201-1204, 1999.

[6] D. Gibbon, U. Gut *Measuring speech rhytm,* Proceedings of the Eurospeech, 2000, Scandinavia, 2001.

[7] W. Jassem *English Stress, accent and Intonation Revisited,* Speech and Language Technology, wyd. PTFon, Poznań, 33 – 50,1999.

[8] D.H. Klatt *Linguistic uses of segmental duration in English: Acoustic and perceptual evidence*, J.Acoust.Soc.Am., vol.59, 5, 1976.

[9] I. Lehiste (1977) *Isochrony reconsidered*, Journal of Phonetics 5, 253 – 265,1997.

[10] R. Ogden, J. Local, P.Carter *Temporal interpretation in ProSynth, a prosodic speech synthesis system*, Proceedings of the XIV th ICPhS, 2, 1059-1062, 1999.

[11] Y. Sagisaka, N. Campbell, N. Higuchi (1997) *Computing Prosody, Computational Models for Processing Spontaneous Speech*, Springer -Verlag, New York, 1997.

[12] J.P. Santen, Ch. Shih *Suprasegmental and segmental timing models in Mandarin Chinese and American English*, JASA 107 (2) 1012-1026, 2000.

ACKNOWLEDGEMENTS

This work was supported by research grant 7T11C00921 from the Polish Committee of Scientific Research.