Polish segmental duration: selected observations based on corpus data

Iloczas głoskowy w języku polskim: obserwacje wybrane w oparciu o dane korpusowe

Katarzyna Klessa

The Institute of Linguistics, Adam Mickiewicz University, Poznań
Adam Mickiewicz University Foundation, Poznań
katarzyna@klessa.pl

ABSTRACT

The paper describes selected observations concerning Polish segmental duration based on data collected in a large speech recognition database. First, a short description of background for the so-far Polish duration investigation is provided with a focus on technology oriented research. Then, a preliminary selection of statistical figures related to Polish segmental duration are presented and the dependence of segmental duration on a selection of interacting factors is discussed.

STRESZCZENIE

W niniejszym artykule zawarto opis wybranych obserwacji odnośnie iloczasu głoskowego w języku polskim, w oparciu o dużą bazę danych, stworzoną na potrzeby rozpoznawania mowy. Najpierw krótko podsumowano dotychczasowe badania nad iloczasem głoskowym dla j. polskiego, ze szczególnym uwzględnieniem badań zorientowanych na zastosowania w technologii mowy. Następnie przedstawiono wstępny wybór danych statystycznych, dotyczących iloczasu głoskowego oraz omówiono wpływ i wzajemne interakcje szeregu wybranych czynników na czas trwania głosek.

1. Polish segmental duration: background

The first publications concerning Polish segmental duration were published in the thirties of the 20th century by Dłuska [1] and Koneczna [2] and included descriptions based on analyses of kymograms. The first results of broader statistical investigation were presented by Lutosława Richter. The subject of one of her works [3] was the analysis of vowel duration in one-syllable and two-syllable logatoms and its dependency on vowel articulation type, the presence of voice in the right direct context, place and manner of articulation of the post-vocalic consonant. The strongest influence was reported for the presence of voice in the following consonant, the weakest for the manner of articulation. In agreement with the author's expectations and also with previous results [4], the duration of particular segments appeared to decrease together with the increase of the length of the whole unit containing these particular segments.

In another paper, the same author [5] compared segmental durations in logatoms and meaningful words, finding a high correlation between the results, and thus concluded that the results obtained for logatoms might be generalized also for words. Later, however, certain important discrepancies were found that showed that such a generalization would not always be appropriate. Summarizing her study on the influence of sound position within an accent unit [6], Richter stated that isochrony was only confirmed for sequencces of logatoms, which was attributed to the test construction and the artificially rhythmical pronunciation. In the same work [6] the influence of stress on segmental duration was confirmed (in agreement with the earlier postulates by Jassem, Dłuska or Koneczna [1, 2, 4]). Three duration classes were distinguished: stressed vowel (the longest), word final vowel (comparably shorter), and a vowel in a pre-stressed syllable or post-stressed syllable but not word-final (the shortest).

In order to finally verify the existence of the isochrony principle in Polish another experiment was conducted by Richter [7], based on a model applied before for Swedish and Dutch [8, 9]. However, the presence of isochrony was confirmed only to a limited extent (i.e. for vowels in selected positions within utterances). The study material for the above experiment consisted of three lists of utterances, each of which was composed of two accent units. The number of syllables within the accent units varied but a similar phonetic structure was maintained within each of the units.

In 1987, Richter published results of testing two alternative methods of rhythm structure modelling [10]. One of the methods was similar to that applied in the 1983 study (a power function expressing the dependency between segmental duration and the number of syllables within an accent unit measured in syllables). The second method was based on a model proposed for English by Jassem et al. In 1981 [11] and used a regression model of the dependency of speech sounds duration on their number within a rhythm unit. Four types of rhythm units were distinguished: a foot (defined as the interval between subsequent accented syllables excluding non-accented syllables in direct post-pausal positions), an accent unit, the ancrusis (the pre-accentual part of the accent unit) and the main part of the accent unit (the accented syllable together with the post-accentual syllables). In the isochrony model, the length of speech segments appeared to be dependent on both the duration of the accent unit and the segment position within the unit. For consonants, the dependency on the number of syllables in the accent unit was not found but it was confirmed that consonant duration depends on their position within the accent unit. The results of regression analysis showed that the most considerable tendency to isochrony was present in the main part of the accent unit, and smallest influence was observed for the sounds in the anacrusis. In comparison to the results for English by Jassem et al. [11] the isochrony in Polish was found to be much less distinctively marked than in English, but it was still consistent enough to make formal descriptions possible.

J. Imiołczyk, I. Nowak, and G. Demenko [12] published a list of factors potentially influencing segmental duration for Polish speech synthesis and corresponding, more detailed segmental duration rules for the vowel /e/. According to the rules, the longest duration among vowels was /a/ while /y/ was the shortest. The unvoiced affricates were treated as the longest among consonants (duration several times as much as the duration of /r/, the shortest consonant). As for the rules related to the properties of the speech segment sequences within the utterance, the longest inter-segment connection durations were assigned to the combinations of /j/ and /y/ with the vowels /a/, /o/, /u/, and the shortest to the combinations of bilabials with /j/. The presence of voice in the direct right neighbourhood was assumed to lengthen segmental duration on vowels, opposite to the presence of a consonant cluster in the right context, which was expected to shorten the preceding vowels. In terms of the manner of articulation of the right context, vowels were expected to be longest when preceding stop consonants and shortest before /r/. The degree of segment shortening within a consonant cluster was assumed to be positively correlated with the number of the cluster components and with the inherent duration of the segment in question (the greatest shortening effect for affricates and no shortening for /r/). The assumed influence of stress was similar to Richter's [6] results (cf. above). As for syllable structure, it was estimated that vowels tend to be longer in pre-pausal syllables when the syllables are open (no consonant in syllable coda). The segmental duration was assumed to be negatively correlated with the number of segments within the rhythmic foot (in agreement with the isochrony principle).

2. Establishing duration modelling features for Polish

Following the previous findings and claims, and based on analyses of Polish synthesis speech corpus a duration model for Polish was designed ([13]; Tables 1, 2 and 3 below) and implemented in BOSS [14]. The duration model has been based on CART duration prediction and includes a total of 57 features from both segmental and suprasegmental levels of the utterance structure [15].

Table 1: Duration modelling features for Polish speech synthesis – properties of phones (bold: highest ranked features in CART cumulative correlation ranking)

Feature types	Description
Inherent properties of the phone in question	**the identity of the current phone** (39 labels, including palatalized [k, g]) (categorical)
	manner of articulation (categorical)
	place of articulation (categorical)
	presence of voice (binary)
	type of the phone in question (consonant or vowel)
Phone properties as related to other units & levels	phone position related to the following pause (float)
	the same preceding phone (binary)
	the same following phone (binary)
	phone position related to consonant clusters (within cluster, preceding/following cluster, no cluster)
	phone position within the syllable (onset, nucleus, coda)
	the same place of articulation as in the direct right neighbourhood within word (binary)
	the same place of articulation as in the direct left neighbourhood within word (binary)
	the same place of articulation across word boundary (binary)
Stress & accent	the presence of stress (binary)
	the presence of phrase accent (binary)

The context information for phone duration is provided for the phone in question and for three adjoining left and right context sounds. The following features are included: the current phone identity, its manner/place of articulation, presence of voice, and type of the phone in question (vowels vs. consonants). The model includes also factors for word and stress information and also the phone position as related to higher level units (syllable, word, phrase and rhythmic foot), and the length and position of higher level units relative to other units of the same and/or other levels of the utterance structure.

The 57-element set of features corresponding to the features shown in Table 2 was used to predict segmental duration with CART (the resulting correlation was 0.8 with RMSE at 15.4, and Error 11.3451). For comparison of the achieved results using varying corpora and feature sets and complete feature ranking cf. [13, 15].

Table 2: Duration modelling features for Polish speech synthesis – properties of nearest context (bold: highest ranked features in CART cumulative correlation ranking)

Feature types	Description
Properties of the 1st, 2nd, 3rd preceding or following context (features similar to the basic properties for the sound in question doubled for the direct left and right context)	**the identity of the first**, second, third **preceding phone** (or: **following phone**)
	manner of articulation of the first, second, **third preceding phone** (or: **following phone**)
	place of articulation of the first, second, third preceding phone (or: following phone)
	presence of voice in the first, second, third preceding phone (or: following phone)
	type of the first, second, third preceding phone (or: following phone)

Table 3: Duration modelling features for Polish speech synthesis – higher level unit organisation (bold: highest ranked features in CART cumulative correlation ranking)

Feature types	Description
Syllable position as related to other units	syllable position within the word (distance from the beginning, in syllables
	syllable position within the word (distance from the word end, in syllables)
	syllable position within the rhythmic foot (in the anacrusis, head or tail)
Rhythmic foot position as related to other units	foot position within the phrase (distance from the phrase beginning, in feet)
	foot position within the phrase (distance from the end of phrase, in feet)
Unit length	word length (in phones)
	the length of the rhythmic foot (in phones)
	syllable length (in phones)
	phrase length
	the length of the whole source utterance

2.1. Improvement of Polish speech synthesis

The perceived quality of the synthesized speech achieved using the above feature set as an input for speech synthesizer was evaluated in a series of perceptual tests, and in the most recent experiments was always assigned the MOS above the grade of 3 on a 5-point scale (i.e. the utterances were intelligible although some acoustic problems were audible, cf. [16]). The duration weighting experiment showed that attributing moderate weights to the applied duration model contributed to better perception of the synthesized speech. It is worth noting that the difference in the perceptual assessment was particularly visible for the conversational sentences set, richer in common structures typical for spoken language. This might suggest that the applied duration model assists unit selection in general, but especially when it comes to synthesizing speech characterized by prosodic structures closer to spoken language. Moreover, since Polish BOSS unit selection in the latest version is based only on the phone level units it appears especially important to deliver information from higher level structures of the utterance in another way, and this task is partly fulfilled by using the multilevel duration feature information.

3. Using a large vocabulary speech corpus for duration analysis

The measurements of statistical significance of the above features and the resulting duration modelling with CART was based mostly on the Polish BOSS corpus (approximately 2 hours of speech by a professional male speaker), and also on a set of recordings of 40 voices (20 male, 20 female) reading a short, 25-sentence phonetically balanced text. The results presented further below in this paper were obtained based on a selection of voices from the *Jurisdic* ASR database.

3.1. The present corpus: *Jurisdic ASR database*

The *Jurisdic* speech database is a large continuous speech database for speech recognition, probably the largest one for the Polish language available at the moment: above 1500 annotated sessions of speakers from 16 regions of Poland, plus another 500 experimental recordings (for more details cf. [17]). A typical scenario for a recording session in the *Jurisdic* database was established according to the *SPEECON* guidelines [18] with necessary adjustments, and is a mixture of semi-spontaneous (controlled dictation) and read speech (in a dictation style). The database includes continuous speech (complex sentences), phrases of varied length, as well as separate words or digits. The database was designed according to the functional requirements of a dictation system for the purposes of courts, police and lawyers' offices, which obviously influences both lexical and syntactical structure of its contents.

Apart from the typical scenarios, the database contains several other types of scenario (original legal texts read or quasi-spontaneously dictated by lawyers). The results presented below in this work are based on a selection of 300 recording sessions from the *Jurisdic* database. The selection included sessions recorded using typical *SPEECON*-standard scenarios. Out of the whole set, 150 sessions were recorded in the Western part of Poland (regions of Poznań, Gorzów, Września, Kalisz, Leszno, Bydgoszcz), and the other 150 recordings come from East of the country (mostly from Białystok, Lublin, Olsztyn).

3.2. Transcription and annotation

All recordings have been first manually labelled on the orthographic level, by a team of students and experts. The phone-level transcription of all the recordings was obtained automatically using the transcriptions stored in the *Speechlabs.ASR* lexical database [19]. The database transcriptions had been generated with an external automatic transcription tool, *Polphone* [20] integrated into the lexical database editor. The quality of the output of the tool has recently been enhanced as a result of work done within an ongoing project (see Acknowledgements). Apart from bug fixes, the recent changes have concerned implementation of more elaborate accent insertion rules and a richer transcription exception list. The automatic phone-level segmentation was performed with *Salian* [21].

4. Segmental duration corpus observations

Figure 1 provides a general view of duration changeability of consonants in contemporary Polish. The scatterplots show the mean values of duration for each of the 300 speakers (above 100 thousand utterances of various length, and 5 million speech sounds), divided into groups depending of the region of origin of the speaker (East and West). The division of the speakers into two regional groups was arbitrary and might be a matter of further discussion. Obviously, a detailed investigation into regional variety would require distinguishing more than just two groups, with a view to find out about the more subtle dialectal differences such as those distinguished in [22] (e.g. including differentiation between Mazovian and Lesser Poland dialects or between Greater Poland and the so-called 'new mixed dialects' of the Western parts of Poland).

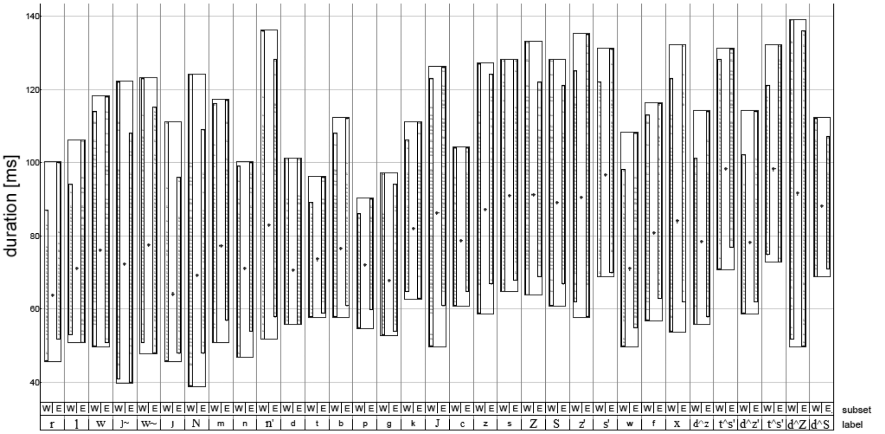

Figure 1: Polish consonant changeability graph. Inside boxes: scatterplot of mean duration values for 150 speakers in the Western (W, left) and 150 in the Eastern (E, right) Poland subsets, dots in the middle of each box stand for mean durations for all speakers for particular consonants.

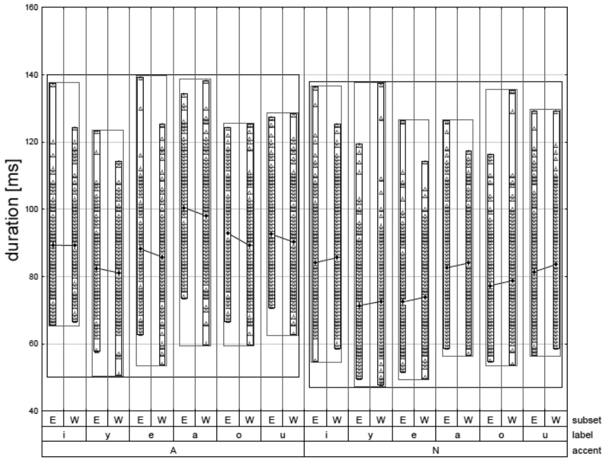

Figure 2: Polish vowel duration changeability graph A- accented, N-non-accented; (inside vowel boxes: scatterplot of mean duration values for 150 speakers in the Eastern (E) and 150 in the Western (W) Poland subsets, dots in the middle of each box (joined with lines) stand for mean durations for all speakers for particular vowels).

The inspection of the results obtained for consonants stays in line with the common claims of rather small regional differentiation of Polish: the durations are quite randomly distributed across the groups. However, certain regularities might be already observed for just the two regional groups while comparing tendencies in vowel lengthening or shortening, depending on the presence of phrase accent (assumed to be placed at the end of a phrase), see Figure 2. This Figure has been produced using a selected portion of the recorded scenario for each of the speakers, namely: four types of read utterances and two types of quasi-spontaneous monologues. The author decided to exclude isolated digits, words, jargon vocabulary etc., in order to keep the more natural and continuous scenario sections for this part. The two types of speech (read utterances vs. quasi-spontaneous) are quite similar in terms of duration changeability, with the quasi-spontaneous utterances having slightly more variability accross speakers in general (regardless of the regional classification).

On average, the durations of accented vowels tend to be slightly higher in the ‘Eastern’ subset than in the ‘Western’ ones (except /i/, where the overall means are practically equal, and the difference is statistically insignificant). On the other hand, the durations of non-accented vowels appear to be a little higher in the West. As far the comparison between accented and non-acented vowels is concern: all mean values for the first are higher than those for the latter, and the analysis of variance showed that these differences are statistically significant. However, it should be emphasized here that this is valid only for the phrase accent. When it comes to word stress, the matter becomes somewhat vague. In a preliminary investigation into the present data (the influence of stress analysed only for the positions in which it does not coincide with phrase accent), a number of speakers demonstrated quite random distributions of duration values and did not seem to confirm any participation of duration in the expression of word stress in Polish.

5. Discussion

Polish segmental duration and its contribution to the perceived speech rhythm has been investigated for quite a long time. The contribution of a number of features has been confirmed, and various configurations of factors are successfully used in speech technology to model segmental duration, however it is quite often the case for researchers to summarize the durational phenomena in Polish with phrases such as: *to a certain extent* (e.g. application of the isochrony principle to Polish as compared to English [7, 11]), *somewhere in between* or even: *neither does it show a behaviour "in between"* (the latter cited after a discussion in [23] as related to the question of the shortcomings of the syllable-timed versus stress-timed languages when it comes to Polish).

The results of duration prediction with Classification and Regression Trees (CART) for speech synthesis and perception tests using the BOSS TTS engine for Polish showed that using a large and complex duration feature set does improve the final quality of the synthesized speech. However, CART and other statistical methods are reported to be biased with certain limitations as related to coping with large numbers of interacting factors (e.g. [24]). Thus, in order to precisely investigate the influence of particular factors as well as their interactions, it appears that perhaps it might be necessary to search for another methods of analysis based on large natural speech corpora, and (maybe first of all) to test the existing hypotheses using representative data.

In this paper, preliminary results of *Jurisdic* corpus mining have been presented for segmental duration of Polish consonants and vowels produced in continuous speech by 300 speakers from Western and Eastern regions of Poland. It was observed that accented vowels tend to be slightly longer for the speakers recorded in the East, and for the unaccented vowels the tendency was reversed, which might suggest that the application of segmental duration as a correlate of accent could be conditioned regionally.

6. Future work

The tasks scheduled for tests and experiments are among others: 1) to look at the results using more sophisticated dialectal divisions; 2) to specify and implement a more sophisticated definition of phrase accent, so that its position would not be limited only to the phrase-final position [25]; 3) to analyse the realisations of rhythm units [7, 11] in the utterances produced by a large number of speakers using various speaking styles and registers (which entails the need to extend the rather formal *Jurisdic* speech material and use more spontaneous recordings). It is planned to continue the segmental duration analyses based on the corpus mining work, starting from a rather general approach towards more detailed search for indications of the most informative data clustering (i.e. not only to test hypotheses but also to attempt to formulate generalizations based on the material).

Acknowledgements

This work is supported from the financial resources for science in the years 2010–2012 as a development project (National Centre for Research and Development).

References

[1] Dłuska, M. 1933. Próba badań nad trwaniem spółgłosek polskich w zależności od brzmienia. In *Slavia Occidentalis* 12.

[2] Koneczna, H. 1934. Studium eksperymentalne artykulacji głosek polskich. In *Prace Filologiczne* 16.

[3] Richter, L. 1973. The duration of Polish vowels. In *Speech Analysis and Synthesis* (Warszawa) 3, 87–115.

[4] Jassem, W. 1962. *Akcent języka polskiego*. Wrocław: Ossolineum.

[5] Richter, L., 1974. Porównanie iloczasu samogłosek polskich wymówionych w logatomach oraz w wyrazach. In *Biuletyn Polskiego Towarzystwa Fonetycznego* 32. 173–178.

[6] Richter, L. 1978. Wpływ pozycji w zestroju akcentowym na czas trwania głosek. In *Lingua Posnaniensia* 21. Poznań.

[7] Richter, L. 1983. *Wstepna Charakterystyka izochronizmu zestrojowego w jezyku polskim. Prace Instytutu Podstawowych Problemów Techniki Polskiej Akademii Nauk* 1983/4.

[8] Lindblom, B. 1968. Temporal organisation of syllable production. In *Speech Transmission Laboratory Quarterly Progress and Status Report* 1 (6).

[9] Noteboom, S. 1972. *Some timing factors in the production and perception of vowels*. Institute for Perception Research. Eindhoven.

[10] Richter, L. 1987. Modelling of the rhythmic structure of utterances in Polish. *Studia Phonetica Posnaniensia* 1, 91–125.

[11] Jassem, W., M. Krzyśko, P. Stolarski. 1981. *Regresyjny model izochronizmu zestrojowego w sygnale mowy. Prace Instytut Podstawowych Problemów Techniki Polskiej Akademii Nauk* Warszawa 1981/33.

[12] Imiołczyk, J., I. Nowak, G. Demenko. 1994. High intelligibility text-to-speech synthesis for Polish. In *Archives of Acoustics*.

[13] Klessa, K. 2006. *Analiza iloczasu głoskowego na potrzeby syntezy mowy polskiej*. Unpublished PhD dissertation, Adam Mickiewicz University, Poznań.

[14] Szymański M., K. Klessa, S. Breuer and G. Demenko. 2010. Polish unit selection speech synthesis with BOSS: extensions and speech corpora. In *International Journal of Speech Technology*, 85–99.

[15] Klessa, K., M. Szymański, S. Breuer and G. Demenko. 2007. Optimization of Polish segmental duration prediction with CART. In *Proceedings of ISCA SSW, Bonn*, 77–80.

[16] Szymański M., K. Klessa, S. Breuer and G. Demenko. 2011. Optimization of unit selection speech synthesis. In *Proceedings of XVIIth International Congress of Phonetic Sciences, Hong Kong*. 1930–1933.

[17] Klessa, K. and G. Demenko. 2009. Structure and Annotation of Polish LVCSR Speech Database. In *Proceedings of Interspeech Conference, September 6–10 2009, Brighton, UK*. 1815–1818.

[18] Fischer, V., F. Diehl, A. Kiessling and K. Marasek. 2000. Specification of Databases – Specification of annotation. SPEECON Deliverable D214.

[19] Klessa, K., M. Karpiński, O. Bałdys and G. Demenko. 2009/2010. Speechlabs ASR. Polish Lexical Database. for Speech Technology: Design and Architecture. *Speech and Language Technology*. 12/13 (Poznań), 191–207.

[20] Demenko, G., Wypych, M., & Baranowska, E. 2003. Implementation of grapheme-to phoneme rules and extended SAMPA alphabet in Polish text-to speech synthesis. *Speech and Language Technology*, Poznań, Vol. 7.

[21] Szymański, M., and S. Grocholewski. 2005. Transcription-based automatic segmentation of speech. In *Proceedings of 2nd Language and Technology Conference, Poznań*. 11–15.

[22] Karaś, H., ed. 2009. Gwary polskie. Przewodnik multimedialny, http://www.gwarypolskie.uw.edu.pl (accessed 15 March 2012).

[23] Wagner, P. 2008. *The rhythm of language and speech: Constraining factors, models, metrics and applications*. Habilitationsschrift, University of Bonn.

[24] Möbius B., and J. van Santen. 1996. Modeling segmental duration in German text-to-speech synthesis. In *Proceedings of the International Conference on Spoken Language Processing, Philadelphia, PA* 4, 2395-2398.

[25] Francuzik, K., M. Karpiński, J. Kleśta and Szalkowska, E. 2004. Nuclear Melody in Polish Semi-Spontaneous and Read Speech: Evidence from Polish Intonational Database 'PoInt', *Studia Phonetica Posnaniensia* 39.