Syllable timing patterns in Polish: results from annotation mining

Dafydd Gibbon [1], *Jolanta Bachan* [2], *Grażyna Demenko* [2]

[1] Fakultät für Linguistik und Literaturwissenschaft, Universität Bielefeld, Germany
[2] Institute of Linguistics, Adam Mickiewicz University, Poznań, Poland

gibbon@uni-bielefeld.de, jolabachan@gmail.com, lin@amu.edu.pl

Abstract

Previous studies of duration variation in syllable constituents have yielded results for Polish which are clear outliers in relation to those for other languages. We report on a study of this issue in the context of TTS development, using a large annotated database. Global and local duration distance measures are applied to phoneme and syllable level units, and generalised iambic and trochaic duration patterns are compared with grammatical structure. The study suggests that Polish is more syllable-timed than previously thought, and that there is tendentially a relationship between syllable duration patterns and word sequences.

Index Terms: Polish, timing, duration, speech synthesis, annotation mining.

1. Introduction

Studies of duration in prosodic typology using read aloud data have shown Polish to be something of an outlier in respect of syllable constituent timing: relatively regular vowel durations, as in syllable-timed languages, but even greater variability in consonantal interval durations than in putative foot-timed languages; cf. [1] and [2]. The syllable constituent parameters concerned are: vocalic-consonantal interval ratio, vocalic and consonantal interval variability [1], length difference between vocalic intervals [2]. The measures are broad-ranging in terms of the languages they investigate, but rely on small data sets. The studies have also been criticised for not treating syllable and foot units [3], and for overstating the relevance of the measures to rhythm modelling [4].

The present study uses a large, richly annotated and rigorously validated corpus, also with read-aloud data, designed for unit-selection speech synthesis, and investigates segment and syllable units, rather than consonantal and vocalic intervals. The approach differs from many signal-based duration modelling approaches for speech synthesis, in that it focusses on sequential patterns in annotations. Clearly care is needed when generalising a single subject study.

We first outline duration determining factors in Polish, then summarise previous results and analyse duration relations at segment and syllable levels, and relate temporal patterns to words. Finally we draw provisional conclusions and point out related ongoing investigations.

2. Duration-determining factors

2.1. Phonotactic factors

It is likely [5] that standard deviations of syllable constituent intervals relate straightforwardly to phoneme counts, and that the results provide rough phonetic correlates, for instance, for the following properties of Polish phonotactics in rhetorically neutral speech [6]:

1. The low variability of vowels in Polish (in rhetorically neutral contexts) may relate to:
   - absence of vowel length contrast,
   - absence of vowel reduction (except for phrase final position, very fast tempo, speaking style, dialect, proximity to consonant clusters).
2. The very high variability of consonants and low proportion of vowels for Polish may relate to:
   - the complexity of Polish onset consonant clusters: ≤5 word initially, medially and finally, and more across word boundaries),
   - proclitic consonantal prepositions (e.g. /v/, /z/).

Polish vowel intervals appear to relate to syllable-timing, but consonantal intervals are highly variable, which may prevent Polish from being syllable-timed. We hypothesise

1. that a Zipf effect (‘longer items are rarer’) counteracts consonantal influence, and
2. that durations of segmental and syllable units in Polish are rather regular.

Polish syllable timing as such has not previously been investigated with the methods applied here. Polish constitutes a relatively clear case for this purpose, since vowels are clearly syllabic and consonants non-syllabic, therefore the distinction between consonantal and vocalic sequences is more clear-cut than in languages such as English.

2.2. Other factors

2.2.1. Linguistic and discourse factors

Several other factors are known to influence duration of syllable constituents [6]. Vowel length may depend, for instance, on vowel type, voicing of neighbouring consonants, position in word and phrase relative to stress and accent, focus, contrast, emphasis, emotion and other rhetorical factors. It is apparently the case that Polish has phrase-final syllable lengthening, and that there is a tendency to foot isochrony [7].

2.2.2. Variable vowel duration in emotional prosody

Informal observations show that in spontaneous Polish speech, in emphatic, emotive, or other rhetorically marked contexts, vowel duration variability is extremely high, in contrast to findings for rhetorically neutral read aloud utterances. A functional reason for this may be that the degree of freedom resulting from lack of phonemic vowel contrast makes this parameter available for other purposes.

Rhetorical variability of vowel duration is not directly relevant to the data types represented in the corpus. It is, however, a factor which needs to be controlled for in empirical studies, and rhetorical differences in data genres (gender, age, register, style) may contribute to the differences pointed out by [2] between their results and those of [1].

3. Procedure

3.1. The data

The data consist of 5 sets of different utterance types designed for unit-selection TTS, read aloud by an adult professional male native speaker of Polish, and richly annotated with boundary and stress types. The total duration is 120 minutes, with 3244 utterances, mean utterance length 2.2 seconds. The read aloud data type was selected in order to reduce dimensionality and to create as near to a clear-case and application-friendly data set as possible. The following example (D1069) shows spelling, English translation, and annotation in modified and enhanced Polish SAMPA:

“To nie jest baryłka, tylko beczka.”
“This is not a small barrel, but a cask.”
`[#to#n'*e#jest#ba.r"yw.ka#t"yl.ko#b&e.tSka]`

(Enhanced diacritic set: # word boundary; . syllable boundary; * , ″ , & accent types. Cf. [8].)

The annotation format was normalised to <label,duration> pairs [9]. For efficiency in processing the large corpus, all operations, including the models described below, were scripted in a Linux environment.

3.2. Smoothness measures

3.2.1. Standard deviation

Several variance-like methods have been used to examine the smoothness of syllable timing (isochrony), e.g. [1], [2] with intervocalic (consonantal) and vocalic intervals, [10], [11] with inter-stress (foot) intervals. Such measures were compared with other measures by [4]. Standard deviation is used in the present study to approximate to these approaches, and the units analysed are segments and syllables, not intervocalic and vocalic intervals, or feet.

3.2.2. PVI

A popular method for studying timing relations is Nolan’s *PVI* model of local evenness of durations. The *PVI* is derived from a standard set distance measure; its application to sequences of speech units is not indisputable, but yields interpretable results. The *PVI* is claimed to measure rhythm.

The *nPVI* averages the normalised distance between adjacent pairs (e.g. segment, syllable), multiplied by 100; normalised distance is the absolute difference between durations of adjacent units, normalised by dividing by the average duration of the units. The *nPVI* values range from 0, i.e. totally even, towards an asymptote of 200, i.e. increasing irregularity; cf. [12].[1] The *rPVI* is a raw distance measure with no normalisation. In [2], *nPVI* is used for vowel intervals and *rPVI* for consonantal intervals.

However, the *nPVI* provides a only necessary and not a sufficient correlate of rhythm. First, it only addresses binary patterns. Second, the absolute difference invalidates the alternation criterion for rhythm: it can easily be verified that the *nPVI* yields the same values for alternating sequences such as 2, 4, 2, 4, ..., geometrical series such as 2, 4, 8, 16, ... or any mixture of these. However, on an interpretation of *nPVI* distance as “local smoothness”, it is a useful measure.

3.2.3. Syllable constituent analyses

The 18 languages treated by [2] include 7 languages treated by [1] (Table 1). Figures from the two sources (separated by “/” in Table 1) show considerable discrepancies for Polish in the standard deviation values, but not in the proportion of vocalic intervals. In the overall variation space bounded by minima and maxima in Table 1, discrepancies affect maxima only.

Table 1: Syllable constituent duration indices for Polish and other languages, summarised from [2].[2]

Parameter	*Polish*		*Min (all)*		*Max (all)*	
% vocalic:	42.3	41	41.1	40.1	55.8	53.1
Intervocalic SD:	71.4	51.4	31.9	35.6	71.4	53.5
Vocalic SD:	44.9	25.1	20.7	25.1	76.4	46.4
Vocalic nPVI:	46.6		27.0		65.8	
Consonantal rPVI:	79.1		40.0		79.1	

The data of [1] and [2] show a low standard deviation for vocalic intervals, comparable with that of syllable-timed languages. The standard deviation of intervocalic intervals is very high (the highest of any of the languages studied), suggesting that Polish may not be a syllable-timed language.

In [3] both the exclusive use of syllable consituents in [2] rather than syllables or feet in analysing timing patterns and also the use of the speech-rate dependent *rPVI* are criticised. In [3], the *nPVI* measure is also applied to syllable and foot units. This approach is followed in the present study, in which the *nPVI* is applied at segment and syllable levels.

3.3. Annotation mining for Polish

3.3.1. Levels of analysis

In order to obtain a more differentiated view of timing in Polish, two levels were analysed:

1. *Phoneme units*, not distinguishing between segment types, because duration gradients also apply within consonantal and vocalic intervals; [1] and [2] imply that segment durations are highly irregular.
2. *Syllable units*, using the clear distinction between the strictly non-syllabic consonants and strictly syllabic vowels of Polish; [1] and [2] imply that there is no clear syllable or foot timing.

3.3.2. Descriptive statistics

Table 2 and Table 3 show results for the 5 data sets.

[1] Division by total length would yield a more conventional asymptote of 100.

[2] In Table 2 of [2] the *nPVI* and *rPVI* columns for Polish have erroneously been reversed. This is evidently a typographical error and has no substantive effect on the authors’ argumentation.

Table 2: Data set: N_{utt} = number of utterances, Dur_{tot} = total duration in min, Dur_{utt} = mean utterance duration in sec).

Set	N_{utt}	Dur_{tot}	Dur_{utt}	*Description*
A	288	13	2.7	Rich in consonant clusters
B	109	3	1.7	Semantically unpredictable
C	669	18	1.6	W. all accented CVC triphones
D	1030	29	1.7	Keyword carrier sentences
E	1148	57	3.0	Very long sentences
All	3244	120	2.2	

Table 3: Analysis of segments and syllables. Tok:Type is the Token:Type ratio; Δt is duration in ms. The minimum measurable duration is 30 ms.

Segments	*A*	*B*	*C*	*D*	*E*	*All*
Tokens	11289	2160	15533	23650	52565	105197
Types	41	40	40	41	41	41
Tok:Type	275	54	388	577	1282	2566
Min Δt	30	30	30	30	30	30
Max Δt	238	263	230	256	310	310
Mean Δt	71	84	70	74	66	70
Median Δt	66	79	65	68	61	65
SD Δt	27	30	26	28	25	26
StdErr Δt	0.25	0.65	0.21	0.18	0.11	0.08
nPVI	39	36	36	35	36	36
Syllables	*A*	*B*	*C*	*D*	*E*	*All*
Tokens	4221	803	6321	9476	20959	41780
Types	1335	506	923	1111	2785	3874
Tok:Type	3	2	7	9	2	11
Min Δt	30	30	30	30	30	30
Max Δt	781	550	585	575	620	781
Mean Δt	195	229	173	186	166	179
Median Δt	182	221	162	173	155	167
SD Δt	71	74	58	70	60	64
StdErr Δt	1.09	2.61	0.73	0.72	0.41	0.31
nPVI	39	40	35	40	37	38

Token and type counts and ratios are a useful indicator of the representativity of the corpus. The large number of syllable types is an indicator of the complexity of Polish phonotactics. The lengths of units vary considerably at the extremes, with expected skew of the duration scale. Whether the range is due to rhetorical factors, or to a personal, gender-specific or dialectal style, has not been determined.

Standard deviation and *nPVI* of both segments and syllables are very constant across data sets, indicating that smaller data sets may provide useful results. Segment variation is relatively large in relation to the mean, which is expected because consonants and vowels were not distinguished. Syllables are very evenly distributed, and the *nPVI* is near the lower (isochrony) end of the range (27...67) reported by [2] for vocalic intervals. It is also consistently about 10 *nPVI* points lower than the *nPVI* for Polish vocalic intervals in [2], possibly due to different phonemic analyses. This may indicate that there is a compensatory effect with longer consonantal intervals occurring with shorter vowel intervals yielding relatively even syllable intervals, but this requires further investigation.

Another consideration is that the durations of consonant clusters were not investigated here, and it is likely that the longer the cluster, the less frequent it is (Zipf effect), and the lower its overall influence on syllable duration. The absence of phonemic vowel length variation may also contribute to the syllabic "smoothness" of rhetorically neutral speech.

Present results therefore indicate more clearly than in previous work first, that Polish may tend towards syllable-timing; second, that vowel duration alone is not necessarily a predictor of syllable duration; third, that compensatory durational adjustment of segments may be involved. Polish seems to be more syllable-timed, for example, than Estonian (syllable *nPVI*=46) and English (syllable *nPVI*=54); cf. [3].

3.4. Structured patterns

3.4.1. Duration chunk patterning

An initial study of the relation between timing units and grammatical units was undertaken, using the same data.

As already noted, the *nPVI*, useful though it is, does not show alternating or rhythmic patterns, contrary to claims in the literature: it shows distance, but does not specify directionality of difference (e.g. longer or shorter), i.e. the rhythmic directionality of iambic (short-long, decelerando) or trochaic (long-short, accelerando). Nor has it been used to investigate the relation of duration patterns to lexico-syntactic structural patterns, for which a concept of directionality (though not necessarily of alternating patterning) is required.

As a first step towards modelling directionality and lexico-syntactic structural patterns, a simple chunking algorithm was used, which items of increasing length (i.e. not necessarily binary) together. Extensions to take other annotation parameters into account are in progress. This simple algorithm promises more structural information than evenness measures, and more direct interpretability than the complex tree construction algorithms introduced by [12]. The core of the algorithm is:

```
create empty accumulator string for labels
for all durations
  append current label to accumulator
  if current duration > next duration
    emit contents of accumulator
    empty accumulator
next duration
```

Labels are strung together until the duration of a label interval is longer than that of the following unit, then stored, and the accumulator is re-set for the next chunk. This is the "iambic" or "decelerando" algorithm.

The converse pattern is generated by the "trochaic" or "accelerando" algorithm, which chunks labels into sequences of decreasing duration. Both algorithms were applied to each of the 5 data sets. Chunks were automatically compared with the annotated (non-phrase-final) word final boundaries to determine the match between chunk ends and word ends. Each utterance was chunked separately and results were averaged (cf. Table 4).

Table 4: Percent iambic:trochaic chunk to word matches for segments and syllables.

Datasets:	*A*	*B*	*C*	*D*	*E*	*All*
N_{utt}	288	109	669	1030	1148	3244
Segments	14:14	22:20	18:14	19:18	17:17	17:17
Syllables	43:36	64:60	47:38	53:47	49:39	50:41

Percentages for all segment chunk types are low: duration patterns inside words are uneven. Syllable chunk matches are a little higher; particularly for the rallentando pattern and in some data sets. Set B (meaningless sentences) has the smallest difference between iambic and trochaic conditions, and set E (long sentences) has the largest. Sets A, C and D show smaller preferences for increasing duration sequences leading up to word final boundaries. In the best case, there is 50% coincidence of duration chunk end and word end. Variation between different data sets is high, perhaps a function of different data types. In view of the small single-speaker data set and the obvious variation between data sets, more detailed analyses are needed which go beyond the scope of the present study.

4. Conclusions and further work

Duration patterns of segments and syllables in a large, richly annotated and rigorously validated single-speaker read corpus of Polish were examined with standard deviation and *nPVI* measures. The study focusses on sequencing rather than classification, and thus differs from the classification and regression tree (CART) type duration models used in speech synthesis [6]. However, the results of the present analysis may enter into the criteria for CART analyses at a later stage.

We have shown that syllable timing in rhetorically neutral Polish speech could tend more towards the syllable isochrony end of the syllable timing scale than has previously been thought. Further, segment sequences turned out to be somewhat even, but the role of consonantal sequence variability in syllable timing thus turns out to be less than previously thought. It remains to be seen whether this means that the more complex consonant clusters of Polish are less frequent than simpler items, whether the lack of a phonemic vowel length contrast in Polish leads to shorter vowels which have more uniform durations and are more comparable on average with consonant durations, and whether there is compensatory duration modfication which supports syllable timing.

Finally, there appears to be a mild tendency for Polish syllable sequencing to be of the generalised iambic (rallentando) type, though results are not conclusive. This needs confirmation from further studies using more annotation categories, on the lines of [13]. The word analyses need to be complemented by analyses based on inter-stress intervals (feet), along the lines of [3], and by investigation of the role of different levels of stress and different boundary types. Comparable results based on these measures are not yet available for other languages, so no typological comparisons can be made at this stage.

5. Acknowledgements

The authors are grateful to many, but particularly to Wiktor Jassem for numerous discussions on prosody. His influence on this study is pervasive. This work was partially carried out under grant R0003502 from the Polish Ministry of Scientific Research and Information Technology and with an EU ERASMUS Student Mobility grant.

6. References

[1] Ramus, Franck, Marina Nespor, & Jaques Mehler,. Correlates of linguistic rhythm in the speech signal. Cognition 72: 1-28, 1999.

[2] Grabe, Esther & Ee Ling Low. Durational Variability in Speech and the Rhythm Class Hypothesis. In Gussenhoven, C. & Warner, N. (eds.), Studies in Laboratory Phonology 7. Cambridge: CUP, 515-546, 2002.

[3] Asu, Eva Liina & Francis Nolan. Estonian rhythm and the Pairwise Variability Index. Proc. Fonetik 2005, Gothenburg

[4] Gibbon, Dafydd & Flaviane Romani Fernandes. Annotation-mining for rhythm model comparison in Brazilian Portuguese. Proc. Interspeech/Eurospeech 2005, 3289-3292.

[5] Hirst, Daniel. Prosodic aspects of speech and language. In K. Brown (ed.) Encyclopaedia of Language and Linguistics. 2nd edition. Oxford, Oxford University Press, Vol. 10:167-178.

[6] Klessa, Katarzyna. Modelowanie iloczasu głoskowego na potrzeby syntezy mowy polskiej. (Modelling the duration of speech sounds for the needs of Polish speech synthesis.). Doctoral thesis, UAM Poznań, 2006.

[7] Demenko, Grażyna. Phrase time structure modelling for speech synthesis purposes. Proc. ICPhS 15, 2441-2444, 2003.

[8] Demenko, Grażyna, Mikołaj Wypych, Emilia Baranowska. Implementation of Grapheme-to-Phoneme Rules and Extended SAMPA Alphabet in Polish Text-to-Speech Synthesis. Speech and Language Technology (7): 79-97, Zakład Graficzny UAM, Poznań, 2003.

[9] Bachan, Jolanta & Dafydd Gibbon. Close copy speech synthesis for speech perception testing. Investigationes Linguisticae (13), Poznań, 9-24, 2006.

[10] Roach, Peter. On the distinction between ‘stress-timed’ and ‘syllable-timed’ languages. In: Crystal, David (ed.), Linguistic Controversies: Essays in Linguistic Theory and Practice, London: Edward Arnold, 73–79, 1982.

[11] Scott, Donia R., Stephen D. Isard, and Bénédicte de Boysson-Bardies. On the measurement of rhythmic irregularity: a reply to Benguerel. Journal of Phonetics 14, 327–330, 1986.

[12] Gibbon, Dafydd. Time types and time trees: prosodic mining and alignmment of temporally annotated data. In: Sudhoff, Stefan, Denisa Lenertová, Roland Meyer, Sandra Pappert, Petra Augurzky, Ina Mleinek, Nicole Richter & Johannes Schließer (eds.), Methods in Empirical Prosody Research. Berlin: Walter de Gruyter, 281-309, 2006.

[13] Hirst, Daniel & Caroline Bouzon. The effect of stress and boundaries on segmental duration in a corpus of authentic speech (British English). Proc. Interspeech 2005, 29-32, 2005.