Analysis of Polish Segmental Duration with CART

Stefan Breuer, Katarzyna Francuzik, Grażyna Demenko

Institute of Communication Sciences
Dept. of Communication, Language and Speech
University of Bonn
Institute of Linguistics, Dept. of Phonetics
Adam Mickiewicz University, Poznań
breuer@ikp.uni-bonn.de, kasia@poczta.com.pl, lin@amu.edu.pl

Abstract

Segmental duration was investigated in a database of Polish read speech (from one male speaker). The material was labeled automatically and then manually verified. The dependence of phone duration on a set of features was verified with the CART algorithm. The duration phenomena were analyzed in relation to syllable, foot and phrase structure. The results showed the need of segmental as well as suprasegmental modeling for the analysis of segmental duration.

1. Introduction

Duration control is a standard requirement that needs to be met to provide the naturalness and intelligibility of the output obtained from speech synthesis systems. Models for predicting duration may be constructed using various techniques: e.g. defining and testing rules, analysis of regression, neural networks or classification and regression trees (CART)[1][2][3][4][5][28]. Generally, the more traditional, rule-based technique follows linguistic cues, and the remaining ones are corpus-based. However, in practice it is often the case that the two kinds of approach overlap. Careful linguistic feature extraction in the stage of data preparation may significantly improve the results of statistical processing. Usually, duration control is considered either on the level of the phone [6] or the syllable [7]. In rule-based concatenation systems, the diphone may also be used as a unit for prediction [2]. Segmental duration may be influenced by both speech rhythm and intonation, which is why they are often considered together, and the relationships between them are explored in search of a model appropriately reflecting the temporal speech structure [8][9].

In this study, the CART technique was used to predict segmental duration on the phone level. The influence of a preliminary set of features on phone duration was examined as well as the possible duration modifications on the level of the syllable, the foot and the phrase.

2. The data and the features

The study material for this paper consisted of almost 50 minutes of recordings read by a male professional native speaker of Polish (the total number of phones was 42082, and after removing pause labels 39104 cases remained). The recordings come from a larger database on Polish segmental duration (in preparation) containing the corpus used here as well as three other corpora (a short text read by 40 speakers, isolated phrases (20 speakers), and a longer, emotionally marked read text (2 speakers). A part of the present corpus was recorded in two speech rates: normal and a faster rate.

2.1. Segmentation and labeling

Initially, the material was labeled automatically using the CreatSeg software (currently developed under the name Salian) [10]. The software enables utterance labeling into phonetic-acoustic segments. The obtained segmentation accuracy is 10 ms. In the second step, the labeling was carefully examined and corrected manually on the basis of visual inspection of spectrograms.

The applied transcription format was a modified Sampa [11] [12][13]. The following sound labels were applied: [a, b, c, d, d^z, d^Z, d^z', e, f, g, i, J, j, k, l, m, n, N, n', o, p, r, s, S, s', t, t^s, t^S, t^s', u, v, w, x, y, z, Z, z', ow~, ew~, aj~, ej~, oj~, @]. The labels [c, J] were added to mark the palatalized allophones of /k, g/. Original SAMPA symbols for the nasalized vowels [e~] and [o~] were abandoned in favor of [ew~], [ow~] before fricative consonants /f, v, s, z, S, Z, x/ and in favor of [e], [ow~] when occurring word finally. The labels [e~] and [o~] before palatal fricatives [s', z'] were replaced by [ej~] or [oj~] respectively. As for the label [aj~], it was introduced to mark the possible variants of realizations of the grapheme sequence: *ań* in a position preceding a fricative as in the word *pański* (typical pronunciation: [p a n' s c i] or an alternative pronunciation variant: [p aj~ s c i] (cf. [13] for more details). Initially, we intended to split each of the five labels into two parts (oral and nasalized), however in our experiments with synthesizing speech sounds of this kind for the above contexts (e.g. [14]) better results were observed when the aforementioned sounds were synthesized from one compound segment rather than from two separate ones. It should be mentioned that in the present material the diphthong labels were used sporadically (only 170 cases). The label [@] referring to the glottal stop was included into the phone set, but its usage was marginal, so statistically it is not important. For more details on the transcription of Polish, see [12][13].

Word stress was marked on the last but one syllable (with a few changes introduced after perceptual verification) which is the norm for Polish. Phrase boundaries were established taking linguistic cues into consideration and then verified on the basis of perceptual evaluation of intonation contours, intensity and pauses. The annotation format applied was the BLF (Boss Label Format) (cf. [15]).

In the preparation stage before the CART analysis, basic statistics were run for the study material using STATISTICA software in order to assess the potential contribution of

particular features in duration modeling and to verify the correctness of segmentation and labeling procedures. These are going to be the subject of more detailed examinations in the near future.

2.2. The features

In the first step of the CART analysis, a set of 52 features was established for preliminary processing. The following list of features was used to define the features:

- Phone identity (43 categories, see 2.1 above)
- Articulation manner (11 categories)
- Articulation place (10 categories)
- Presence of voice (2 categories)
- Sound type (3 categories)
- Pre/Post-pausal phone position (3 categories)
- The same/different place of articulation in the preceding/following phones (2 categories: the same or different articulation place)
- The same/different neighboring phone in the preceding/following phones (2 categories: the same or a different phone)
- Position relative to consonant clusters (4 categories)
- Position within the syllable structure (3 categories: position in onset or nucleus, or coda)
- Syllable position within the word (syllable distance from the beginning/end of the word counted in syllables) (float)
- Syllable position within the foot structure (3 categories)
- Foot position within the phrase (foot distance from the beginning/end of the phrase counted in feet) (numerical value)
- Sound position within the phrase (3 categories)
- Stress (3 categories)
- Word length (number of phones) (numerical value)
- Speech rate (2 categories: normal and fast)

The sound classes determined by the features 'Articulation manner', 'Articulation place', 'Presence of voice', and 'Sound type' were defined both for the given phone and for its preceding and following context. The context was verified for the phones directly adjacent to the sound in question, for the post-following and pre-preceding ones and also for the 3rd phone before and after the sound. For the feature 'Articulation place' the possible durational contribution of the following categories was checked with the CART analysis: bilabial, palatal, dental, labio-dental, velar, alveolar, labio-velar, back vowel, front vowel, palatalized vowel. The sound class 'Articulation manner' was divided into categories as follows: fricative, affricate, nasal, w, j, r, l, vowel, nasalized vowel, and stop. For the 'Sound type' class, three categories were used: vowel, consonant, and compound vowel. The 'Pre/post-pausal position' feature also had three categories: pre-pausal phone position, post-pausal phone position and phone position non-adjacent to any pause. For 'Position relative to consonant clusters', three categories were considered: phone position within a cluster of more than two consonants, phone position directly following such a cluster and phone position with no direct neighborhood of any consonant cluster. For the feature 'Syllable position within the foot structure' the notion of the foot and anacrusis (see e.g. [16]) was accepted. We understand the foot as the interval between two subsequent stressed syllables. Anacruses may occur utterance-initially and comprise of syllable(s) preceding the first stressed syllable of the utterance. Three categories were distinguished in this class: the syllable position in the foot's head (the first syllable of the foot, i.e. the stressed syllable), in its tail (one of the unstressed syllables following the stressed one) or in the anacrusis. In due course we intend to check the effect of the foot in more detail (especially its length and relations with syntactic structure). For the class 'Stress', three categories were taken into account: the last word stress of a phrase (nuclear accent), word stress (pre-nuclear), and no stress. The sound position within the phrase could be initial, medial or final.

2.3. Remarks on feature selection

The above list of features was established on the basis of other studies and previous data for Polish e.g. [17][18][9]. Only part of them proved to be significant both in the preliminary variance analysis and by using the CART method of analysis. The size of the corpus (almost 50 minutes of recordings of read speech) seems to be representative. It is hard to expect great improvement of the results by enlarging the dataset.

What seems to be more important is to ameliorate the factorization scheme. It seems that suprasegmental features, , especially speech rhythm should be investigated in more detail. Including speech rhythm information is essential for duration modeling, but so far solutions for implementing the information for the purposes of automatic analyses have not been satisfactory.

In many languages speech rhythm is related to the phenomenon of isochrony. In so-called stress-timed languages, the rhythm units are assumed to have a relatively constant length, regardless of the number of syllables ([19][20]). In these languages (e.g. English, German, Dutch, Polish) speech sounds are shortened to a certain degree along with the lengthening of the rhythm unit. Jassem [21] presented an analysis of isochrony useful for practical applications. However, it should be stated that the notion of isochrony has been controversial for some of the languages and it is often subject to further investigation. One of the more interesting solutions to the problem of duration modeling for TTS purposes is the Prosynth model based on phonological theory (Ogden et.al.[22]), which is a syllable-based model. Simplified isochrony models based on vowel duration analyses have also been developed, such as the PVI - Pairwise Variability Index (e.g. [8][23]). So far, however, none of the existing utterance structure models has permitted unconstrained speech rate and speech rhythm control.

3. Results

To obtain our results, we used the CART implementation *wagon* which is part of the Edinburgh Speech Tools [24].

We performed the training using 95% of the data for training and leaving out 5% for testing. This yielded an RMSE of 25.86 ms and a mean correlation of 0.62.[1]

We then used the *stepwise* option of *wagon* to estimate the contribution of our features. This time, the complete dataset was used. The resulting RMSE equaled 25.60 ms and the mean correlation was 0.63. The obtained feature ranking is presented in Table 1.

In the first column, the feature names are given and the second column gives the cumulative correlation values between the

[1] See [24] and [28] for details on these parameters.

observed and predicted duration obtained in the CART modeling process.

Table 1: *Feature ranking.*

Feature	Corr.
Phone identity	0.4088
Following phone identity	0.5425
Preceding phone identity	0.5626
Articulation manner of the 3rd following phone	0.5764
Syllable position within the foot	0.5847
Articulation manner of the preceding phone	0.5900
Foot position within the phrase (distance to the right phrase boundary)	0.5936
Word length in phones[1]	0.5967
Articulation manner (the phone in question)	0.6010
Presence of voice (the phone in question)	0.6066
Speech rate	0.6102
Presence of voice in the following phone	0.6132
Articulation manner of the post-following phone	0.6151
Articulation manner of the pre-preceding phone	0.6167
Sound type (the phone in question)	0.6180
Articulation manner of the following phone	0.6208
Presence of voice in the preceding phone	0.6217
Articulation place in the following phone	0.6225
Foot position within the phrase (distance to the left phrase boundary)	0.6229
Sound position within the phrase	0.6237
Syllable position within the word (distance from the beginning of the word)	0.6242
Articulation place in the post-following phone	0.6248
Articulation place in the preceding phone	0.6251
Presence of voice in the pre-preceding phone	0.6254
Presence of voice in the post-following phone	0.6256
Sound type of the preceding phone	0.6259
Sound type of the post-following phone	0.6262
Articulation place of the pre-preceding phone	0.6264
Syllable position within the word (distance from the end of the word)	0.6265
Presence of voice in the 3rd following phone	0.6266
Sound type of the following phone	0.6268
Position relative to consonant clusters	0.6269
Pre-preceding phone identity	0.6269
Sound type of 3rd following phone	0.6270
Articulation place of 3rd following phone	0.6270
Post-following phone identity	0.6271
Stress	0.6271

The results show, not surprisingly, that the most important feature is the identity of the phone itself. The next ones are those connected with the preceding and following context identity, syllable position within the foot, and the articulation properties (articulation place, sound type, the presence of voice, and articulation manner) of the given segment and its context. The contribution of the features connected with stress, segment boundaries and structure appeared to be much smaller. However, as the *wagon* manual states, the ranking provided by stepwise training should not be used to derive a general order of importance of the selected features.

The above results were compared to the results obtained from the CART analysis run for a smaller and more homogeneous dataset of Polish read speech using a more limited number of features. For a dataset of 11795 vectors of 8 parameters, an RMSE of 19.3 ms and mean correlation of 0.677 was obtained, which is not substantially different from the outcome for the present material.

In order to check if the results would improve, a series of tests were run using various features sets applied to chosen subsets of the database. The general tendency observed was that removing subsets from the corpus and limiting features number deteriorated the overall performance of the CART processing. The effect of excluding corpus subsets recorded in two speech rates and the subsets containing isolated phrases is shown as an example in Table 2 below. The table presents also the results obtained from only the largest subset of the database, which appeared to achieve the highest mean correlation value and the lowest RMSE. This set is described as "emotionally marked". While it is not surprising that a more uniform set results in a better durational prediction, this could also be a hint that our feature set does not sufficiently capture the variation induced by speaking style and affect.

Table 2: *CART results for various data subsets.*

Test	RMSE [ms]	Mean Correl.	n
3 sets removed	31.32	0.53	19823
2 sets removed	28.86	0.57	26302
1 corpus subset	20.08	0.69	6586

The highest correlation for the identity of the segment in question was reported by many researchers for various languages (cf., [26] [27]**Fehler! Verweisquelle konnte nicht gefunden werden.**). The order of the subsequent features varies across different studies, however, in most of them the identity of the neighboring segments is mentioned among the top ones (e.g.,[5][26]), similar to the present results. Further features observed to be significant by other researchers are those related to phrase structure and rhythmic properties of the utterance. The latter were not sufficiently explored for the present study, which may explain the relatively low mean correlation (0.62 as compared with 0.79 for Czech [5], 0.75 for Hindi [26], or 0.73 for Korean [27]. The RMSE was relatively good in the present experiment (25,86 ms) as compared with the same parameter in the example studies mentioned above (20.3, 27.1 and 26.5 ms respectively).

4. Discussion

Difficulties related to automatic duration modeling were discussed in detail in many studies (cf., [7][8]) and the fundamental problems in this matter have already been solved for read speech. Duration prediction by means of learning techniques (especially with the popular CART algorithm) enables direct implementation in speech synthesis providing relatively correct output. However, feature revision and adding new features of the suprasegmental level seems necessary, especially with respect to synthesizing spontaneous and expressive speech.

[1] Some of the feature values for word length were miscalculated, so the actual contribution may be different.

In our study, several features of the segmental level showed 60% correlation, and adding other features (most of them of the segmental level) did not significantly improve the results. What appears especially important now, is a careful selection of features influencing duration both on the segmental and the suprasegmental level within the intonational-rhythmic phrase structure.

5. Conclusions

In spite of a relatively low correlation between observed and predicted durations, the feature ranking order was in agreement with results reported earlier (see section 3 above). Nonetheless, further statistical analyses and testing are required to verify the feature selection made mostly on the basis of linguistic presumptions. For the purposes of duration modeling of various speech styles (e.g. with attitude or emotion), adding new features, especially those describing the internal structure of feet and a more sophisticated analysis of the suprasegmental level might be necessary.

It would also be desirable to check our factorization scheme with learning techniques other than CART (e.g. neural nets, regression analysis, rough sets) and rules.

6. References

[1] Klatt, D. H. 1976. *Linguistic uses of segmental duration in English; Acoustic and perceptual evidence*. JASA 59 (5), pp. 1208 - 1221.

[2] Olaszy, G. 2002. *Predicting Hungarian Sound Duration for Continuous Speech*. Acta Linguistica Hungarica, vol. 49 (3-4), pp. 321 - 345.

[3] Riedi, M.P. 1998. *Controlling segmental duration in speech synthesis systems*. PhD thesis, TIK-Schriftenreihe (26), ETH Zürich.

[4] Vainio, M. Altosaar, T; Karlajainen, M; Aulanko, R; Werner, S. 1999. *Neural network models for Finnish prosody*. Proceedings of ICPhS'99, California.

[5] Batusek, R. 2002. *A Duration Model for Czech Text-to-Speech Synthesis*, Proceedings of Speech Prosody, Aix-en Provence.

[6] Moebius, B., van Santen, J.P.H. 1996. *Modeling segmental duration in German text-to-speech synthesis,* Proceedings of the International Conference on Spoken Language Processing (Philadelphia, PA). (4) pp. 2395-2398.

[7] Campbell, N. 1992. *Multi-level timing in speech*, PhD Thesis, University of Sussex (Exp. Psychol): Brighton, UK.

[8] Gibbon, D. 2005. *Cognition, Perception and Measurement. On the Modeling of Rhythm*, SASR Kraków.

[9] Demenko, G. 2002. *Modelowanie czasowej struktury frazy na potrzeby syntezy mowy.*, Speech and Language Technology (6), Poznań.

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

[11] Wells, J. 1996. The SAMPA homepage: http:/www.phon.ucl.ac.uk/home/sampa/home.htm

[12] Jassem, W. 2003. *Illustrations of the IPA: Polish*, Journal of the International Phonetic Association (33/1), pp. 103-107.

[13] Demenko, G., Wypych, M., Baranowska, E. 2003 *Implementation of Grapheme to Phoneme Rule and Extended Sampa Alphabet In Polish Text-to-Speech Synthesis*. Speech and Language Technology (7), pp. 79-95. Poznań.

[14] Francuzik, K, 2003. *Kontynuacja prac nad konwerterem TTS dla języka polskiego opartym na technologii Profivox* (*A continuation of work on TTS converter for Polish based on Profivox technology*) (©TUB-TTT 1995-2003); KBN Report.

[15] Breuer, S., Wagner, P., Abresch, J., Bröggelwirth, J., Rohde, H., Stöber, K. 2005. *Bonn Open Synthesis System (BOSS) 3 Documentation and User Manual.* http://www.ikp.uni-bonn.de/boss/BOSS_Documentation.pdf

[16] Jassem, W., Krzyśko, M., Stolarski, P., 1981. *Regresyjny model izochronizmu zestrojowego w sygnale mowy*, IPPT PAN, Warszawa.

[17] Frąckowak-Richter, L.1973. *The duration of Polish Vowels, Speech Analysis and Synthesis* (3), 87-116, PWN: Warszawa.

[18] Frąckowak-Richter, L. 1987. *Modelling the rhythmic structure of utterances in Polish*, Studia Phonetica Posnaniensia (1), 91-125.

[19] Lehiste, I. 1977. *Isochrony reconsidered*, Journal of Phonetics 5, 253 - 265, 1997.

[20] Sagisaka Y., Campbell N., Higuchi N. 1997. *Synthesizing Spontaneous Speech in Computing Prosody* In: Sagisaka Y., Campbell N., Higuchi N. eds., *Computing Prosody, Computational Models for Processing Spontaneous Speech.* Springer-Verlag New York, Inc., 165 - 185.

[21] Jassem, W. 1999. *English Stress, accent and Intonation Revisited,* Speech and Language Technology (3), 33 - 50 Poznań.

[22] Ogden, R., Local, J., Carter, P., 1999. *Temporal interpretation in ProSynth, a prosodic speech synthesis system*, Proceedings of the 14th ICPhS,2, 1059-1062.

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

[24] King, S., Black, A.W., Taylor, P., Caley, R., Clark, R. 2003. *Edinburgh Speech Tools. System Documentation Edition 1.2, for 1.2.3 24th Jan 2003.* http://www.cstr.ed.ac.uk/projects/speech_tools/manual-1.2.0/

[25] Demenko, G.,1999. *Analiza cech suprasegmentalnych języka polskiego na potrzeby technologii mowy,* Wydawnictwo Naukowe UAM,, Poznań.

[26] Krishna, N.S, Murthy, H.A. 2004. *Duration Modeling of Indian Languages Hindi and Telugu*. Proceedings of 5th ISCA SSW Pittsburgh.

[27] Chung, H., Huckvale, M.A. 2001. *Linguistic Factors Affecting Timing in Korean with Application to Speech Synthesis*. Proceedings of Eurospeech, Scandinavia.

[28] Breiman, L. et al., 1984. *Classification and Regression Trees*. Belmont, Wadsworth.

ACKNOWLEDGEMENTS:

This work was supported by Alexander von Humboldt Foundation and by KBN (Project no. 2 H01D 003 24).