Optimization of Polish Segmental Duration Prediction with CART

Katarzyna Klessa, Marcin Szymański**, Stefan Breuer***, Grażyna Demenko**

Institute of Linguistics, Dept. of Phonetics
Adam Mickiewicz University, Poznań, Poland*
Poznań University of Technology, Poland**
Institute of Communication Sciences
University of Bonn, Germany***
katarzyna.klessa@amu.edu.pl, marcin.szymanski@cs.put.poznan.pl,
breuer@ifk.uni-bonn.de, lin@amu.edu.pl

Abstract

This paper describes results of the investigation of Polish segmental duration for the purpose of speech synthesis. The experiment is a continuation of the previous work of the same authors [1] aiming at improving the outcome of the duration prediction mechanism to enhance the overall quality of synthesized speech.

1.Introduction

Duration prediction models for speech synthesis range from the more traditional, rule-based techniques to trainable, corpus-based techniques. Nowadays, it is often the case that the two approaches overlap and careful linguistic feature extraction usually is an important stage preceding the actual statistical processing [2, 3, 4, 5, 6] The unit that should be regarded as the base for segmental duration modeling is also a subject of discussion. Most frequently, phone is used as the unit, however there are also other proposals e.g. Campbell's syllable-based model [7]. Linguistic knowledge may be used not only in the data preparation process but also in the modeling process itself which is postulated and tested for various languages by Van Santen's sum-of-product models e.g. [8, 9, 10]. In the present study, phones were the base units for prediction, though the influence of other units was considered. In our experiment, the influence of two sets of factors on phone duration was investigated. Then we compared the present results with the ones we obtained from similar tests done with a smaller corpus of data [1]. Both experiments were performed using BOSS technology [11].

2. Corpora and annotation procedure

For the most recent analyses, we used a corpus of two hours of continuous speech read by a male professional speaker. The results were then compared to those obtained from a corpus of almost 50 minutes of speech produced by the same speaker [1].

- The 2-hour corpus contains utterances prepared especially for the database to provide coverage for the most frequent Polish triphones (most importantly CVC triphones in the contexts of sonorants), as well as the most frequent diphones, and consonant clusters and the 600 most common Polish words.
- The texts in the 50-minute database consist of fragments of prose, newspaper articles, short dialogs, (some of them having a potential emotional load), and a list of railway enquiry entries.

Both corpora were annotated according to the same labeling procedure. First, the recordings were labeled automatically with "Salian" [12]. (the segmentation accuracy of the software is 10 [ms]). In the second step, the labels were corrected manually based on visual inspection of spectrograms. Both corpora were revised by the same group of experts following the same guidelines. The 50-minutes corpus was corrected mainly using Wavesurfer, and for the other corpus "Annotation Editor", a tool designed specifically for the purposes of the speech synthesis project for the Polish language (see acknowledgments) was used. SAMPA for Polish was used as the transcription alphabet with the following modifications:

- Palatalized variants of [k][g] were added to the label set and marked with: [c][J]
- Labels for the Polish nasalized vowels i. e. [e~][o~] were removed and instead sequences of vowels and nasalized consonants [w] or [j] marked with [w~][j~] were used

The second modification will be subject of further investigation in the future as there are reasons to suspect that better results for synthesis could be obtained by connecting the oral and nasal component into compound items [13]. Syllable boundaries were inserted automatically as well as word stress labels. Word stress was assigned to the penultimate syllable, which is its most common position in Polish. Afterwards, the actual placement of stress was manually verified and corrected, if needed. Phrase boundaries were established according to linguistic cues and then verified on the basis of perceptual evaluation of intonation contours, intensity and pauses.

3. Features for training CART

Initially, the list of features for duration prediction included the following information:

- Sound (which particular phone is the phone in question)
- Sound's properties. The following features were included as sound properties: the manner of articulation, the place of articulation, the presence of voice, the type of sound (consonant or vowel)
- Properties of the preceding and of the following context. The properties were exactly the same as those listed above as the properties of the sound in question. A 7-element frame was used as the context information, i. e. the same properties were used as features for three preceding and for three following phones as well as for the phone in question.
- Position within the higher unit of speech organization structure. Feature space included: position of the phone in question relative to pause; the distance of the syllable containing the phone to the left and right word boundary; the position of the syllable within the foot (in the anacrusis, head or tail of the foot), the position of the foot within the intonation phrase. Position of the sound within syllable structure (onset, nucleus, coda).
- Identical neighborhood. The information if the phone in question occurred in the neighborhood of an identical phone in the directly preceding or following context within the same word.
- The position of sound relative to consonant clusters (within cluster, before or after cluster or with no cluster in the direct neighborhood).
- Word length and foot length.
- Word stress and sentence stress.

4. Results

To obtain our results, we used the CART implementation "wagon" [14]. The set of features corresponding to the properties listed in the previous paragraph was used to predict segmental duration with the 50-minute database and with the 2-hour database. The results are shown in the first two columns of Table 1. Each time, the results were obtained with 5 % held out data.

Table 1: Comparison of CART results for two corpora

50-min Corpus	2-h Corpus	2-h Corpus & Features Modified
RMSE 19.1973 **Correlation**: 0.7284 Mean (abs) **Error** 14.1092 (13.0182).	**RMSE** 15.5178 **Correlation** : 0.8047 Mean (abs) **Error** 11.4132 (10.5139)	**RMSE** 15.4010 **Correlation**: 0.8080 Mean (abs) **Error** 11.3451 (10.4154)

As can be seen, the results for the 2-hour corpus are significantly better. The difference in RMSE (the root mean squared error) appeared to be more than 4 milliseconds, the overall mean correlation also improved from almost 0.73 to 0.8. In search for further improvement of the prediction the list of features was extended by the following new items:

- The same or different place of articulation of the phone in question and the phone in its direct left or right context
- The same place of articulation across word boundary. The information if the phone in question occurred in a neighborhood of a phone with the same place of articulation across word boundary
- Syllable length, phrase length, and the length of the whole source utterance

After including this information into the feature space, there was another improvement of the results, yet it was less substantial than the one obtained after switching to a bigger and better controlled database in terms of the coverage of phonetic-acoustic properties of the read speech.

The next step was to check how the contribution of particular features to the overall result obtained by the whole feature set was influenced after adding new features. In order to get the information, we used the stepwise option of "wagon" [14]. With the stepwise option enabled, results are expressed as cumulative correlation.

The number of features in the two experiments differed by six items (51 features in the initial test, and 57 features after adding the new features). As it appeared, the order of features in order of their contribution was very similar in both of the two experiments as far as the most contributing features are concerned. The first difference was observed on the 12th position in the ranking. In the experiment with the modified feature list the 12th feature was one of the new features: the length of the utterance. As for the other new features, the phrase length was on the 16th position, the syllable length on the 18th position, and the information of the same or different place of articulation across word boundaries was on the next position, immediately followed by the feature "same or different place of articulation" for the right context.

Table 2 shows the 15 most contributive features in the test performed with various number of features. In the first two columns the results for the two-hour database are presented for experiment with the two different feature lists. The last column shows the similar results for the 50-minute database that were obtained using the shorter feature list.

Despite the differences between the corpora and various number of features taken into account four features out of the first fifteen in the rankings appear exactly on the same position in each of the three sets (marked in bold).

Another observation is the that in the results for the 50-minute database, the first feature in the ranking is the "phone in question", which is consistent with the results reported for many similar studies for various languages e. g. [6, 15, 16]. The second feature in terms of importance was the immediate right context. For the two-hour database the order appeared to be reverse: the phone in question was placed on the second position just behind the "right context" feature. First, we thought one of the possible reasons might be the fact that Polish diphthongs were treated as two separate units in the 2-hour database while in the 50-minute database they were regarded as compounds. In order to check if that was the case an additional test using stepwise option of "wagon" was run. This time, two parts of diphthongs were again joined into compound units, however in the resulting order of features the right context was still more contributive than the phone in question.

Table 2: Feature ranking comparison (stepwise) – the most important features, cumulative correlation.

2-h corpus features modified (Dataset of 98835 vectors of 57)	2-h corpus (Dataset of 98835 vectors of 51)	50-min corpus (Dataset of 98835 vectors of 51)
1. Right context: 0.5062 2. Phone in question: 0.7004 **3. Left context: 0.7375** 4. **Foot distance to the right phrase boundary: 0.7613** 5. Articulation manner in the 3rd right context: 0.7703 6. Articulation manner in the left context: 0.7749 7. Nuclear stress: 0.7787 **8. Presence of voice (the phone in question): 0.7846** **9. Articulation manner in the right context: 0.7893** 10. Articulation manner (the phone in question): 0.7934 11. Articulation manner of the 2nd left context: 0.7963 12. Utterance length: 07893 13. Presence of voice in the left context: 0.8003 14. Word length: 0.8025 15. Articulation manner of the 2nd right context: 0.8038	1. Right context: 0.5062 2. Phone in question: 0.7004 **3. Left context: 0.7375** 4. **Foot distance to the right phrase boundary: 0.7613** 5. Articulation manner of the 3rd right context: 0.7703 6. Articulation manner in the left context: 0.7749 7. Nuclear stress: 0.7787 **8. Presence of voice (the phone in question): 0.7846** **9. Articulation manner in the right context: 0.7893** 10. Articulation manner (the phone in question): 0.7934 11. Articulation manner of the 2nd left context: 0.7963 12. Articulation manner of the 2nd right context: 0.7982 13. Syllable distance from the beginning of the word: 0.7995 14. Sound type in the right context: 0.8003 15. Presence of voice in the left context: 0.8011	1. Phone in question: 0.4774 2. Right context: 0.6396 **3. Left context 0.6625** **4. Foot distance to the right phrase boundary 0.6793** 5. Articulation manner in the left context 0.6859 6. * Syllable position within the foot 0.6919 7. * Articulation place (the phone in question) 0.6959 **8. Presence of voice (the phone in question) 0.7033** **9. Articulation manner in the right context 0.7098** 10. Articulation manner in the 3rd right context: 0.7146 11. Articulation place in the right context 0.7168 12. Syllable distance from the beginning of the word: 0.7189 13. Articulation manner in the 2nd left context: 0.7208 14. Articulation place in the 2nd right context 0.7219 15. Phone position in the syllable onset, nucleus or coda: 0.7231

It should be noticed that in the classification and regression tree the top-most rule generated for the 2-hour database was "CRIGHT is sil" (i.e. right context is silence). Due to the fact that one of the sub-bases of the 2-hour database consisted of short, 3-4-word phrases, the stronger influence of the following context might possibly be explained by prepausal lengthening effect.

Most features in the first fifteen positions occur in all three tests with two exceptions: the features "syllable position within the foot" and "Articulation place (the phone in question) " appear on the 6th and 7th position in the test for the 50-minute corpus, these two are marked with a star.

Additionally, two more experiments were performed. First, the "new" set of features was used to test a corpus composed of all data, i.e. both the 50-minute and the 2-hour database. The results appeared to be slightly better than for the 50-minute database but worse than those for the 2-hour database. The numbers were as follows:

- no heldout: RMSE 16.3432, Correlation 0.7878, Mean (absolute) Error 11.5945 (11.5182)
- 5% heldout data: RMSE 16.3977 Correlation is 0.7862 Mean (abs) Error 11.6520 (11.5377).

For the second experiment, fifty minutes of recordings were randomly selected from the 2-hour database to compare the outcome of the two corpora using a similar amount of speech data. The results were characterized by the following values of correlation and errors:

- no heldout: RMSE 15.9093 Correlation is 0.7929

  Mean (abs) Error 11.7416 (10.7352)
- 5% heldout data: RMSE 15.9644 Correlation is 0.7912 Mean (abs) Error 11.7652 (10.7909)

The above values are an improvement as compared both to the 50-minute database and to the results obtained for the combined databases, however they are slightly worse than those for the 2-hour database. The latter observation seems to be an obvious effect of enlarging the speech corpus. The deterioration of the results after combining both corpora might be due to differences between the types of texts recorded in the two databases. The speaker tended to accelerate while reading longer texts as compared to short separate phrases even though he was supposed to keep the same rate for all the recordings. However this relation requires further investigation.

5. Conclusions

Correlation and RMSE improved substantially when we used the larger corpus containing sentences prepared to cover the most frequent clusters, diphones and triphones as the training data set. The modification of the list of features provided a further (slight) improvement of the results. The comparison of the results obtained with the two feature sets (Table 2) shows that the first difference in the order of contribution appears on the 12th position (per 51 or 57 items), the first eleven items are ordered identically in the two tests.

This stability of feature order, together with the improvements in correlation and RMSE suggests that our choice of features comprises the chief linguistic and phonetic determinants of segmental duration. The problem that needs further examination is the fact that in the recent tests the feature "phone in question" was the second most contributive feature and not the first one which seems to be the obvious expected result and was always the case in the previous experiments of the authors as well as reported in other studies. The values of correlation and the RMSE after the modification of the corpus and the list of features for duration prediction provided comparably good results. The next step in order to obtain further improvement should be a closer investigation into the cost function in the unit selection algorithm, which will be performed in the near future, as well as more detailed analyses of the statistical relevance of the results obtained for correlation and RMSE

6. Acknowledgements

This paper was supported by Alexander von Humboldt Foundation and by Polish Scientific Committee (KBN Project no. 2 H01D 003 24) .

7. References

[1] Breuer, S., Francuzik, K., Demenko, G., Szymański, M. *Analysis of Polish Duration with CART*, Proceedings of Speech Prosody, Dresden, 2006.

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

[3] Olaszy, G., Predicting Hungarian Sound Durations for Continous Speech. Acta Linguistica Hungarica, Budapest, vol. 49 (3-4), 2002, str. 321-345.

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

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

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

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

[8] V. Santen, J.P.H., Quantitative Modeling of Segmental Duration. [in:] Proceedings of Human Language Technology Conference, Princeton, New Jersey, 1993, str. 323-328.

[9] V. Son, R.J.J.H., V. Santen, J.P.H., Strong interaction between factors influencing consonant duration. [in:] Proceedings of Eurospeech '97, Rhodos, 1997.

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

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

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

[13] Demenko, G., Wypych, M., Baranowska, E. 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, 2003.

[14] King, S., Black, A.W., Taylor, P., Caley, R., Clark, R., Edinburgh Speech Tools. System Documentation Edition 1.2, for 1.2.3 24th Jan 2003.

[15] Krishna, N.S., Murthy, H.A., Duration Modeling of Indian Languages Hindi and Telugu. [in:] Proceedings of 5th ISCA Speech Synthesis Workshop, 2004.

[16] Chung, H., Huckvale, M., Linguistic factors affecting timing in Korean with application to speech synthesis. [in:] Proceedings of Eurospeech 2001, Scandinavia. http://www.tsi.enst.fr/~chollet/Biblio/Congres/Audio/Eurospeech01/CDROM/papers/page815.pdf