OPTIMIZATION OF UNIT SELECTION SPEECH SYNTHESIS

M. Szymański[a], *K. Klessa*[a], *S. Breuer*[b] *& G. Demenko*[a]

[a]Poznań Supercomputing and Networking Center; [b]Adam Mickiewicz University, Poznań, Poland
marcin.szymanski@speechlabs.pl; klessa@amu.edu.pl; stefan.breuer@gmx.net;
grazyna.demenko@speechlabs.pl

ABSTRACT

This paper reports on the improvement of Polish speech synthesis obtained by applying new techniques to BOSS (The Bonn Open Synthesis System) for Polish. In order to enhance the system's performance a variety of set-ups for the cost function, types of units used for concatenation (uniform vs. non-uniform unit selection) and the corpus alignment were tested. Three configurations for segment duration weights were chosen and tested with a mean opinion score perception test to investigate the impact of the applied segmental duration model on the perceived speech quality.

Keywords: speech segmentation, duration models, speech synthesis, unit selection, Polish

1. INTRODUCTION

The perceived naturalness of synthesized speech depends on the supply of appropriate units as well as their combinations which should appropriately reflect coarticulation effects and other between-unit phenomena. The huge number of possible connections between units in natural speech has been a subject of concern for synthesis quality [11], since it unavoidably results in unnatural distortions at concatenation points, even if a TTS (text-to-speech) system works generally well. A related problem is the high occurrence of rare events, the so called LNRE problem (Large Number of Rare Events) which remains to pose a challenge both in the acoustic inventory design and acoustic modeling [10]. A compromise between the database's size and sufficient coverage of unit connections can be reached by optimizing the contents of the database (e.g. using "greedy" set covering algorithms [12]) and also by manipulating the size of the units used for unit selection. Non-uniform unit selection has been reported to result in a good quality of synthesized speech for many languages e.g. [6, 9] and the possibility of selecting longer concatenation units is expected to result in a smaller number of glitches and a more natural sound. However, for morphologically complex languages, like Polish (or Turkish, Arabic etc.) it is especially challenging to optimize text corpora in a way to obtain a satisfactory quality of synthesis using higher level concatenation units. A huge number of inflected forms would need to be supplied to obtain a comparable number of directly usable units [9]. Apart from the unit size, a large share of the attained speech quality can be attributed to the selection preferences set by cost functions and penalties. Thus, even when only one type of units "takes part" in the selection the selection might be influenced by constraints from different level structures.

The corpora used for the Polish BOSS and the details of the speech corpus annotation and duration modeling have already been described in several publications [4, 8]. The present version of the acoustic corpus contains 2670 sound files (115 min. of speech). The performance of the speech synthesizer has recently been tested using: fully automatic, semi-automatic and fully manual signal segmentation [4]. The results of the segmentation experiments showed that the synthesis based on fully manual segmentation was perceived as insignificantly better than the one obtained with the fully automatic method, while the performance of the synthesizer using the semi-automatic segmentation method gave the worst results as compared to the other two. In the experiments reported in this paper the automatic alignment method was applied [13]. Section 2 summarises the improvements of the Polish version of the BOSS synthesizer obtained by modifying the unit types and cost function weights. Section 3 shows the positive contribution of the segmental duration model. In Section 4 the results are discussed and concluded.

2. UNIT SELECTION AND COST FUNCTIONS ADJUSTMENT

The latest version of the Polish BOSS has been developed testing a variety of set-ups for the cost function, manipulating the influence of the phrase concatenation units boundary type and

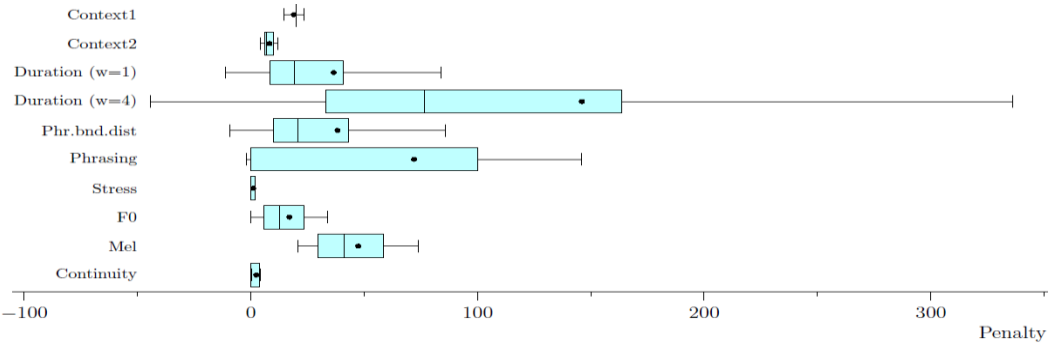
experimenting with the size of (uniform vs. non-uniform unit selection). While previous versions of Polish BOSS followed the top-down approach using words, syllables and phones with pre-selection criteria optimised for a directory enquiries application [1], the present setup uses the phone level exclusively. However, a comparably strong preference for longer units is retained by using the connectedness criterion which supersedes all other transition costs. The pre-selection constraints put on phones to enter the actual cost function consist of a single set including left and right phone context, phrase boundary type and lexical stress. If no units match the requirements, all available phones of the requested type are regarded as candidates.

The unit costs applied in the present version of Polish BOSS consist of the following features listed by: **Duration**: the absolute difference between the CART-predicted segment duration and the candidate unit duration (in ms); **Stress**: a penalty for the mismatch between the predicted and the actual stress value; **Phrasing**: costs for the discrepancy of phrase position and phrase type between target and candidate as given by the intonation phrase and phrase boundary strength features (more in [2, 4]); **Phr. bnd. dist**: a penalty for the normalised distance difference of target and candidate from the phrase boundary; **Context1/2**: costs for phonetic differences between target and candidate in terms of the sound class and manner/place of articulation of the preceding/following context.

In the present implementation, two features are considered by the transition cost function if the units are not connected in the source utterance (no Continuity): the Euclidean Mel Frequency Cepstrum distance between the left segment's right edge and the right segment's left edge, the absolute F0 difference, analogously.

Figure 1 shows a boxplot of a typical distribution of unit and transition costs for the above features. The weights for each cost term were adjusted manually, using preference tests by expert participants to assess whether an improvement has been achieved. The choice for pure phone-based unit selection was made in the same way.

Figure 1: A boxplot of a typical distribution of weighted unit and transition costs for the features (whiskers denote one standard deviation below and above the mean). For an explanation of the features cf. Sec. 2. Two variants of the duration cost correspond to two distinct setups used for the experiments.



2.1.1. Perception tests results

The above unit selection set-up was used to generate a set of 74 sentences. The whole set was composed of three subsets of utterances:

- **Common** (25 sentences and phrases created especially for the purpose, mostly using the top frequent vocabulary items from a large vocabulary newspaper frequency list);
- **Conversation** (25 typical Polish conversation phrases, dialogue phrases, short expressions);
- **Natural** (reference set: 24 original recordings of the speaker reading short sentences).

The utterances were presented to 29 subjects students of the Institute of Linguistics. Each of the subjects listened to the samples individually, via headphones, in a quiet room. The MOS (Mean Opinion Score) task was to assign a score to each utterance on a nine-point scale, from 1 (worst) to 5 (excellent) with an 0.5 interval.

The overall MOS result for the synthesized speech (together for the subset **Common** and **Conversation**) was 3.39 with a standard deviation of 0.89, while for the **Natural** speech recordings set the mean was 4.6 with a smaller standard deviation of 0.49. All results were statistically significant (p-value < 0.0001). The latest experiments focusing on duration weighting for the system's duration model are presented in Section 3.

3. SEGMENTAL DURATION WEIGHTING

3.1. Segmental duration model

The duration model used in Polish BOSS is based on CART duration prediction and includes a total of 57 features from both segmental and suprasegmental levels of the utterance structure ([7, 8]). The context information for phone duration is provided for the phone in question and for three adjoining left and right context sounds. The included features are: the current phone identity, its manner/place of articulation, presence of voice, and type of the phone in question (vowels vs. consonants). Moreover, the model includes factors for word and stress information and also the phone position as related to higher level units, namely: syllable, word, phrase and rhythmic foot (mainly 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 above properties was used to predict segmental duration with CART, the resulting correlation was 0.8 (with RMSE at 15.4, and Error 11.3451 (cf. [8]).

3.2. Experiment design and procedure

The text material used in the duration weighting experiment included 45 sentences selected from the **Conversation** (a random selection of 22 sentences) and **Common** (a random selection of 23 sentences) sets enlisted in Sec. *2.2.1*. Three synthesizer configurations were used to observe the impact of the duration model on the perceptual assessment of the synthesized speech:

- duration weight set to 0 - the influence of the duration model set up to **zero** value, thus neglecting information from the duration model.
- duration weight set to 1 - which was expected to influence the output in a **moderate** way
- duration weights set to 4 - presumed **strong influence** on the unit selection.

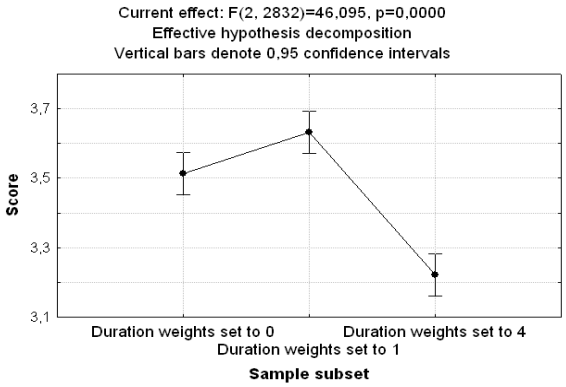
The above configurations were chosen after preliminary tests using a range of various weights. The weights above the value 4 were excluded in order to avoid overriding the join cost function and also costs related to phrases types and boundaries (cf. *2.2.1*). The latter should preserve their impact because of the fact that they represent the F0 but also segmental duration information (phrase final lengthening effect etc.). The choice of only three weights was dictated by practical reasons: to limit the resulting number of sentences necessary to perform a valid MOS test.

The same 45 sentences were synthesized with each of the above duration weight configurations. Then, an MOS test was carried out to observe the perceptual assessment of the samples (nine-scale, like in *2.2.1*). 145 utterances were presented to the subjects in a random order (45 for each of the 3 duration weights configurations, plus 10 additional control samples). 21 listeners took part in the experiment (students or employees of linguistic faculties). The samples were presented via headphones to each person individually, in a quiet room. The testing time ranged from 25 to 40 min. per person (no time constraints were imposed).

3.3. Perception tests results

As a result of the perception MOS test a total of 2835 scores were assigned to the synthesized samples (each of 21 subjects assessed the three variants of 45 sentences). The differences in means between the three sample sets appeared to be statistically significant (in ANOVA tests). As shown in Figure 2, the mean scores were the best for the sample set generated using the configuration with duration weights set to 1 (overall MOS of 3.63, standard deviation: 0.94).

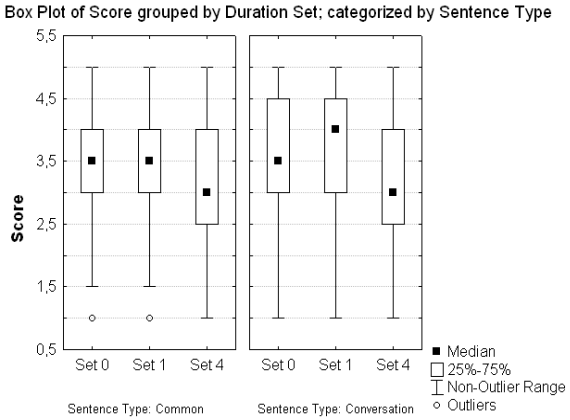
Figure 2: MOS results for three configurations of the synthesizer: duration weights set to 0, to 1, to 4.



The mean for the scores obtained with duration weights set to zero (i.e. ignoring the influence of the duration model) equaled 3.51 (Std. Dev. of 0.93), whereas assigning a higher weight caused deterioration of the mean score (3.22, Std. Dev. 0.98). All listeners showed the same tendency in scoring the samples except from only one person who judged the output of a part of zero weight configuration results slightly better than the ones moderately influenced by the duration model. The differences in mean scores related to the sentence type (Conversation vs. Common) also appeared to be statistically significant, and were especially

visible for the moderate duration weight configuration (Set 1), where the Conversation phrases were rated as better on average than Common sentences (cf. Figure 3). Considering a larger prosodic variability provided by the Conversation sentence set as well as a more "spoken" character of the sentences, this observation suggests that using duration weighting in a moderate way contributed most considerably to a better perception of the synthesized speech. The median values for the 0 and 4 duration setups were very similar for the two types of sentences.

Figure 3: Mean scores grouped by duration weight setup (Set 0, 1, 4) and categorized by sentence type (left: Common, right: Conversation).



4. DISCUSSION AND FINAL REMARKS

The perceived quality of the synthesized speech in all the above experiments for various text types was always assigned the MOS above the grade of 3 (i.e. the utterances were intelligible although some acoustic problems were audible). In the perception tests carried out with the previous version of the synthesizer [1, 3] the overall results of all MOS tests were regularly below grade 3 (an overall MOS result of 2,46 for synthesis based on automatically segmented speech). The primary explanation for the improvement, apart from bug fixes in the cost functions, is the fact that in the former study, the perception tests were carried out using samples generated by non-uniform unit selection, while in the present experiments, phones were the concatenated units with only a single pre-selection step to limit the set of candidates. Significant time was also spent on manually tuning the weights and improving the balance between unit and transition costs (cf. also [2]).

The duration weighting experiment shows that attributing moderate weights to the duration model contributes to better perception of the synthesized speech. The difference in the perceptual assessment was particularly visible for the **Conversation** sentence set, richer in structures typical for spoken language. This suggests 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 (compare [5]). Since Polish BOSS unit selection is now based only on the phone level units it appears important to deliver information from higher level structures of the utterance, and this task is partly fulfilled by using the multilevel duration feature information.

5. ACKNOWLEDGEMENTS

This project is supported by The Polish Ministry of Sciences and Higher Education (project no OR00006707).

6. REFERENCES

[1] Breuer, S., Abresch, J. 2003. Unit selection speech synthesis for a directory enquiries service. *Proc. ICPhS* Barcelona.

[2] Breuer, S., Hess, W. 2010. The Bonn open synthesis system 3. *Int. J. of Speech Technology* Springer, 13(2), 75-84.

[3] Demenko, G., Bachan, J., Möbius, B., Klessa, K., Szymański, M., Grocholewski, G. 2008. Development and evaluation of Polish speech corpus for unit selection speech synthesis systems, *Proc. Interspeech* Brisbane.

[4] Demenko, G., Klessa, K., Szymański, M. Breuer, S., Hess, W. 2010. Polish unit selection speech synthesis with BOSS: extensions and speech corpora. *Int. J. of Speech Technology* Springer, 13(2), 85-99.

[5] Hirose, K. 2008. Speech prosody in spoken language technologies, *Journ. of Sign. Processing* 12(1), 7-16.

[6] King, S., Portele, T., Höfer, F. 1997. Speech synthesis using non-uniform units in the Verbmobil project. *Proc. Eurospeech* Rhodes, 2, 569-572.

[7] Klessa, K., 2006. *Modelowanie Iloczasu Głoskowego na Potrzeby Syntezy Mowy Polskiej*. (*Segmental Duration Modelling for Polish TTS*). Unpublished PhD dissertation, Adam Mickiewicz University, Poznań.

[8] Klessa, K., Szymański, M., Breuer, S., Demenko, G., 2007. Optimization of Polish segmental duration prediction with CART. *Proc. 6th ISCA Workshop on Speech Synthesis* Bonn.

[9] Möbius, B. 1998. Word and syllable models for German text-to-speech synthesis. *Proc. 3rd Internat. Workshop on Speech Synthesis* Jenolan Caves, 59-64.

[10] Möbius, B. 2001. Rare events and closed domains: two delicate concepts in speech synthesis. *The 4th ISCA ITRW on Speech Synthesis* Perthshire.

[11] van Santen, J.P.H. 1993. Exploring N-way tables with Sums-of-Product models. *J. Mathematical Psychology* 37(3), 327-371.

[12] van Santen, J P.H., Buchsbaum A.L. 1997. Methods for optimal text selection. *Proc. Eurospeech* Rhodes.

[13] Szymański, M., Grocholewski, S., 2005. Transcription based automatic segmentation of speech, *Proc. 2nd Language and Technology Conf.* Poznań.