BULLETIN OF THE POLISH ACADEMY OF SCIENCES
TECHNICAL SCIENCES, Vol. 58, No. 3, 2010
DOI: 10.2478/v10175-010-0035-1

Implementation of Polish speech synthesis for the BOSS system

G. DEMENKO[1]*, B. MÖBIUS[2], and K. KLESSA[1]

[1] Institute of Linguistics, Department of Phonetics, Adam Mickiewicz University, 4 Niepodległości Ave., 61-874 Poznań, Poland
[2] Institute of Natural Language Processing, University of Stuttgart, 7 Keplerstraße, 70174 Stuttgart, Deutschland
Institute of Communication Sciences, University of Bonn, 164 Römerstr., 53117 Bonn

Abstract. The Bonn Open Synthesis System (BOSS) is an open-source software for the unit selection speech synthesis that has been used for the generation of high-quality German and Dutch speech. This article presents ongoing research and development aimed at adapting BOSS to the Polish language. In the first section, the origins and workings of the unit selection method for speech synthesis are explained. Section two details the structure of the Polish corpus and its segmental and prosodic annotation. The subsequent sections focus on the implementation of Polish TTS modules in the BOSS architecture (duration prediction and cost function) and the steps involved in preparing a new speech corpus for BOSS.

Key words: speech synthesis, text-to-speech (TTS), unit selection, duration prediction.

1. Introduction

The key idea of corpus-based synthesis is to select at run-time from a large recorded speech database the longest available strings of phonetic segments that match a sequence of speech sounds representing the target sentence. Current unit selection approaches mostly use segments [1–3] or sub-segmental units such as half-phones [4, 5] or demiphones [6] as the basic unit. If units larger than segments are available, the number of concatenations as well as the need for signal processing can be reduced. The frequency of unit concatenations in diphone synthesis (one concatenation point per phone) has been argued to contribute to the perceived lack of naturalness of synthetic speech. In a speech database comprising several hours of recordings, it is likely that a target utterance may be produced by a small number of units each of which is considerably longer than a segment or a diphone (Fig. 1).

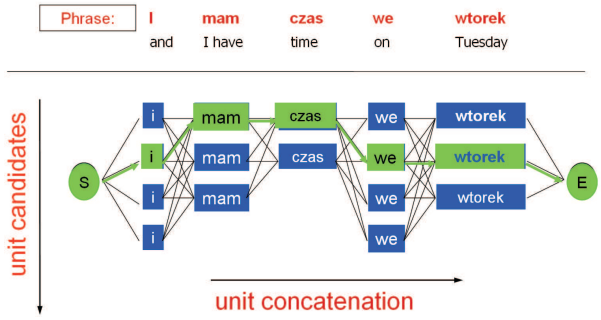

Fig. 1. The target phrase 'I mam czas we wtorek' is synthesized by concatenating word-sized acoustic units. The units are selected from a set of candidates available in the recorded speech database. Selection criteria include the goodness of match of a candidate to the target specification (e.g, finding in the database a 'wtorek' that is stressed and utterance-final) and the smoothness of joining adjacent units (e.g. finding a 'we' and a 'wtorek' that can be concatenated without audible discontinuities)

Defining the optimal speech database for unit selection is a crucial, yet difficult, task in building a speech synthesis system. A well-designed speech corpus has a strong impact on the quality of the synthesized speech. It is now generally accepted that in order to benefit from long acoustic units, a judicious selection or even design of the text materials to be recorded is required. The database should cover all relevant acoustic realizations of phonemes, a point made already by Iwahashi and Sagisaka [7]. However, the enormous combinatorics of features and parameters in language and speech imposes restrictions on the attainable synthetic speech quality, as no corpus can completely cover the set of features required to produce natural sounding speech [8, 9].

Speech synthesis systems are based on machine learning techniques and rely heavily on training with speech material representative of a specific task. The quality of the synthesized speech depends on the text type and synthesis domain: intonation is very natural for restricted domain, e.g. news or weather forecast, and prosodically table speech (read or dictated texts) which is distinguished by quite flat intonation, stable voice quality and easily predictable duration of the speech units. Ideally, the speech segments should cover all phonetic variations, all prosodic variations, and all speaking styles. Due to the limited speech material to be recorded per speaker the focus has to be on the coverage of phonetic and prosodic variations which means that the speaking style should be quite uniform over the domains chosen. In order to meet the requirements concerning the coverage of segmental and suprasegmental features, the size of databases for speech technology purposes is expected to be substantial, e.g. according to ECESS guidelines [10] the overall duration of the recorded speech signals for speech synthesis database should be approximately ten hours.

Criteria for defining the structure of the speech corpus interact with unit selection criteria. A large-scale evaluation is

*e-mail: grazyna.demenko@speechlabs.pl

required to establish the optimal combination of TTS modules and unit selection algorithms.

The Blizzard Challenge aims to compare research techniques for corpus-based synthesis using the same corpus data [11]. Synthesis voice quality is assessed by listeners on the basis of a prescribed set of test sentences. The initiative of the European Center of Excellence for Speech Synthesis [10] attempts to evaluate not only entire TTS systems but also TTS components.

The BOSS TTS system [12–15] is an open source architecture for concatenative speech synthesis, especially for unit selection. BOSS was originally developed for German but the latest version [13] has seen significant changes to software design and architecture that makes it easily extensible to be used in a multilingual context. Several of the system components have been generalized to accommodate other languages, and TTS development for Polish has served as a testbed for the language-independent applicability of the BOSS architecture. The Polish unit selection corpus is described in the following section. The implementation of Polish modules for duration prediction and cost functions for the BOSS system is discussed in Sec. 3, and the results of the system evaluation are reported in Sec. 4, of this paper.

2. Polish Speech Corpus

2.1. Corpus contents and structure. The problem of constructing an effective low redundant database for flexible concatenative speech synthesis has not been solved satisfactorily either for Polish or any other language. We have decided to use various speech units from different mixed databases as follows:

- Base A: Phrases with most frequent consonant structures. Polish language has a number of difficult consonant clusters. 367 consonant clusters of various types were used.
- Base B: All Polish diphones produced in 114 grammatically correct but semantically nonsense phrases.
- Base C: Phrases with CVC triphones (in non-sonorant voiced context and with various intonation patterns). 676 phrases were recorded for triphone coverage.
- Base D: Phrases with CVC triphones (in sonorant context and with various intonation patterns). The length of the 1923 phrases varied from 6 to 14 syllables to provide coverage of suprasegmental structures (the fundamental frequency of recorded phrases varied from 80 Hz to180 Hz).
- Base E: Utterances with 6000 most frequent Polish vocabulary items. 2320 sentences constructed by students of the Institute of Linguistics at the University of Poznań.
- Base TEXT: Continuous text read as whole paragraphs (not separated into sentences on the stage of recording). 15 minutes of prose and newspaper articles.

The entire linguistic material was read by a professional radio speaker during several recording sessions, supervised by an expert in phonetics. The speech errors were corrected online during the recording sessions. Finally the entire recorded material was perceptually verified by another expert.

2.2. Phonetic labeling. The computer coding conventions were drawn up in SAMPA for Polish [16] with revisions and extensions and in the IPA alphabet [17]. Two sets of characters were precisely defined for the exact GTP mapping for the Polish language – an input set of characters and an output phonetic/phonemic alphabet [18]. An inventory of 39 phonemes was employed for broad transcription and a set of 87 allophones was established for the narrow transcription of Polish. Apart from the phone labels enlisted in the above table the symbol "$p" was used to mark a pause, "#" was used for word boundaries. Two additional labels were included: "@" to mark a centralized vowel sound (schwa) and "?" for glottal stop. Formally, glottal stop is not included in the inventory of Polish phones, however speakers tend to produce it at the beginning of vowels after a pause.

SALIAN software has been developed for the automatic segmentation of speech. Its features include: calculating segment (usually phoneme) boundaries based on phonetic transcription, context-dependent phoneme duration models, considering "forced" transition points for semi–automatic segmentation, accepting triphone statistical models trained with HTK tools, tools for duration models calculation, orthographic-to-phonetic conversion, evaluation of decision trees to synthesis unseen triphones, accepting wave or MFCC files (plus several label formats) as input, posterior triphone-to-monophone conversion (for more details see [19]).

2.3. Suprasegmental annotation. The goal of the text analysis component is to convert the input text into a phonological description consisting of a phoneme chain associated with some sort of prosodic and accentual description. The BOSS annotation system requires information about segmental and suprasegmental structure. General intonation theory for Polish is not much different from English or German. The intonational phrase which is determined by the optional pre-nuclear intonation and the obligatory nuclear intonation is assumed to be the largest unit. The intonational phrase is determined by the optional pre-nuclear intonation and the obligatory nuclear intonation. The pre-nuclear as well as the nuclear intonation structure is determined by accentual groups, which carry the secondary real accent or the primary real accent.

The automatically phonetically labeled speech database was annotated for suprasegmental features by four experts on the basis of perceptual and acoustic analyses of the speech signals. On the phrase level annotation of sentence and intonation type was provided. On the syllable level pitch accent types have been marked. On the acoustic level, pitch accents are determined by pitch variations occurring on the successive vowels/syllables and pitch relations between syllables. Pitch accent type annotation can be complex because it may include combinations of many acoustic features (e.g. pitch movement direction, range of the pitch change, pitch peak position, cf. Figs. 2 and 3).

With a view to simplifying the annotation of the pitch accents only two dimensions were considered: the pitch movement direction and its position with respect to accented syllable boundaries (Fig. 4).

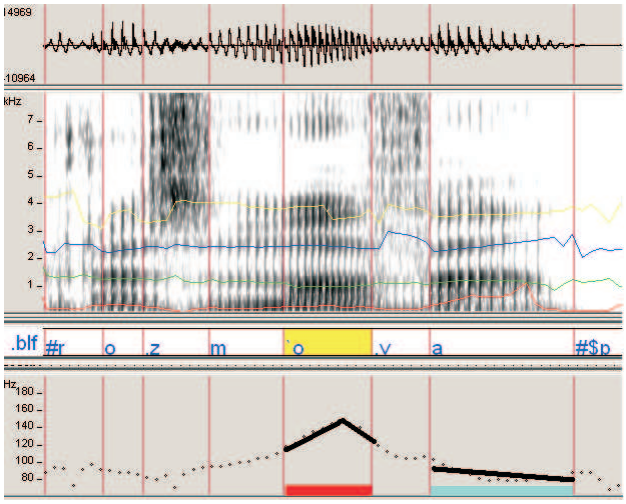

Fig. 2. The temporal alignment of a pitch accent with respect to the syllable boundaries is crucial: The peak of the accent on 'rozmowa' is too late, conveying the (undesired) presence of irony

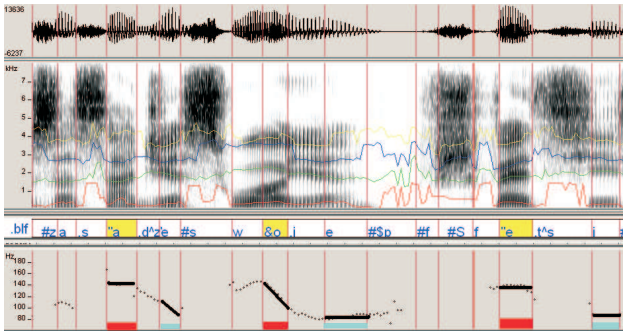

Fig. 3. Selecting units with the appropriate type of pitch accent and temporal alignment contributes to the generation of natural sounding prosody. Here: a sequence of a falling accent with the actual fall on the post-tonal syllable, a falling accent with the fall on the accented syllable, and a falling accent produced by a combination of level tones on the accented syllable (high) and the post-accented syllable (low)

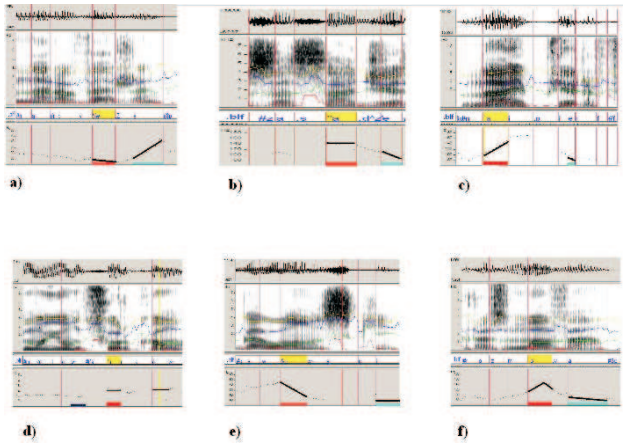

Fig. 4. Pitch accents inventory: a) pitch movement with rising intonation R (on the post-accented syllable: LH) b) falling intonation F (on the post-accented syllables: HL) c) rising intonation on the accented syllable d) level intonation e) falling intonation on the accented syllable f) rising-falling intonation on the accented syllable. Accented syllables are highlighted

The resulting inventory of pitch accent labels include: two labels reflecting pitch movement direction i.e. falling intonation (HL) and rising intonation (LH). In both cases the movement is realized on the post-accented syllable and the maximum/minimum occurs on the accented syllable. Another three labels also reflect the pitch movement direction (falling, rising and level), but the pitch movement is fully realized on the accented syllable. Level accent is realized by duration. Special label describes rising-falling intonation on accented syllable (RF).

For prosody modeling, only fundamental types of suprasegmental structures were distinguished, such as word and phrase accent placement or phrase boundary type according to the BOSS synthesis system format.

Annotation Editor software was created for suprasegmental annotation and also for manual correction of SALIAN's automatic segmentation. The programme supports simultaneous processing of text files, BLF files and spectrographic analyses of the respective sound files (via *Wavesurfer* [20] engine ran from inside of Annotation Editor as if in a plugin mode).

3. Implementation of Polish TTS modules in BOSS

Two Polish modules have been implemented for BOSS so far [21, 22]: the duration prediction module and the cost functions module. In BOSS, cost functions may be effective on both nodes and arcs (representing speech units and concatenations, respectively) of the network of candidate units. Currently, the node cost function applied in the Polish version of BOSS consists of the following components: the absolute difference between the CART-predicted segment duration and the candidate unit duration, the boolean difference between predicted and actual stress value, multiplied by 10, the discrepancy regarding phrase type (question or statement, raising or falling intonation) and phrase location within a sentence (final or comma-terminated), multiplied by 20. In the most recent implementation, two features are considered by the transition cost function: the Euclidean MFCC distance between the left segment right edge and the right segment left edge, the absolute F0 difference, analogously (currently only for phone segments).

The auditory experiments suggest that relocation of the syllable within the phrase should be particularly penalised. Several experiments to predict segmental duration with CART were carried out, using various sub-corpora of the speech database. The best obtained results (the overall correlation of 0.8) were reported in Klessa et al. [21].

Some of the most important factors affecting the temporal structure of Polish (among others as phone type, type of adjoing phones, type of consonant context following the vowel, type of the consonantal cluster, position of syllable in the utterances) have been analyzed within recently carried out research based on a larger database (50 utterances coming from 40 speakers). The detailed analysis showed the importance of rhythm modeling. Phone duration is negatively correlated with the number of syllables co-occuring in a rhythmic foot. Statistical duration models become very useful for different

languages. The model developed for Polish utilizes a neural network to map the relation between phonotatic features and normalized durational values. The correlation between predicted and observed phoneme duration values was relatively high: 83% (fully connected feed forward neural network with Levenberg Marquardt training algorithm).

The present corpus enabled a more comprehensive duration investigation since it contains a variety of texts ranging from short phrases, through longer and more complex sentences up to continuous text, both of rather formal and informal, expressive style. Thus, it became possible to observe the relations between segmental duration and factors both from the segmental and suprasegmental level. The first step of the duration analysis was focused on the distributions, means, and variances of the duration as a variable dependent on a presumed set of modifying factors. In the second step, the usefulness of a set of 57 modifying factors for duration prediction was assessed by means of the Classification and Regression Trees (CART) algorithm [23]. The results support the claim that the duration of speech sounds may be modified by the influence of segmental and suprasegmental features as well as by their combination. The following set of features was taken into account for duration prediction:

- The properties of the sound in question: the information which particular phone is the phone in question, its manner and place of articulation, the presence of voice, the type of sound (consonant or vowel).
- The properties of the preceding and of the following context. The properties were exactly the same as those listed above for the sound in question. In CART analyses 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.
- The position within a higher unit of speech organization structure. (syllable, word, phrase).
- Information about the direct neighborhood of the phone in question (within and across word boundary, relative to properties of adjacent sounds or sound clusters).
- Word length and foot length.
- Syllable length, phrase length, and the length of the whole source utterance.
- Word stress and phrase accent.
- Several experiments to predict segmental duration with CART were carried out, using various sub-corpora of the speech database.

The sound classes determined by the features 'Manner of articulation', 'Place of articulation', 'Presence of voice', and 'Type of sound' 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 (Fig. 5).

For the feature 'Place of articulation' the possible durational contribution of the following categories was checked with the CART analysis: bilabial, palatal, dental, labiodental, velar alveolar, labio-velar, back vowel, front vowel, palatalized vowel. The sound class 'Manner of articulation' was divided into categories as follows: fricative, affricate, nasal, w, j, r, l, vowel, nasalized vowel, and stop. For the 'Type of sound' 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 'Consonant clusters', four categories were considered: phone position within a cluster of more than two consonants, phone position directly preceding/following a cluster and phone position with no direct neighborhood of a cluster. The feature 'Syllable position within the foot structure' was observed as either syllable position in the foot's head or tail or in anacrusis. For the class 'Stress', three categories were taken into account: nuclear accent (the last word stress of a phrase), prenuclear stress, no stress. The sound position within the phrase could be initial, medial or final.

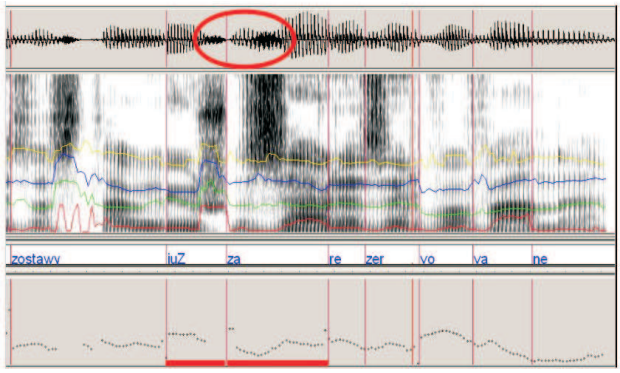

Fig. 5. Selecting units based on matching feature specifications of both the target and the adjacent units can prevent poor concatenations such as the one between the fricatives [Z] and [z] (marked by the red ellipse)

4. Evaluation of speech synthesis quality

The best results of synthesis have been obtained in domain synthesis for train information, because the linguistic structure of this database was carefully prepared. The synthesized speech has a good segmental and rich suprasegmental structure (Figs. 6 and 7).

The utterance: 'Pociąg pośpieszny do Krotoszyna przez Malbork oraz Bydgoszcz wjedzie wyjątkowo na szósty peron na dworcu zachodnim', (Eng. 'The fast train to Krotoszyn through Malbork and Bydogoszcz will be arriving today only at platform number six at the western station'), was built from words, syllables, phonemes (Fig. 6).

Figure 7 shows the example of the synthesized utterance: 'Ta ruda panienka jest szwagierką Marylki', (Eng. 'That redheaded young lady is Marylka's sister-in-law') with linguistic structures not contained in the database used used for domain synthesis for train information. The utterance was built from syllables and phonemes. The segmental features of this synthesized utterance were acceptable, the intonation was not very differentiated, because the suprasegmental structure of database was not representative enough.

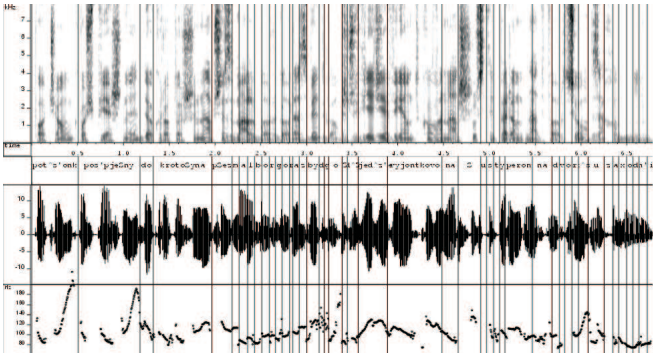

Fig. 6. Spectrogram of the synthesized utterance: 'Pociąg pospieszny do Krotoszyna przez Malbork oraz Bydgoszcz wjedzie wyjątkowo na szósty peron na dworcu zachodnim'

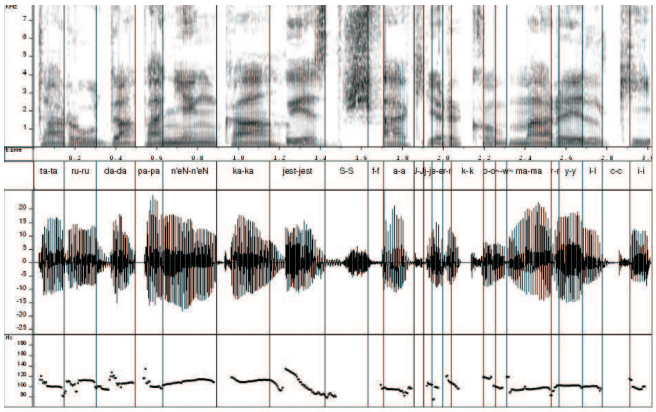

Fig. 7. Spectrogram of the synthesized utterance: 'Ta ruda panienka jest szwagierką Marylki'

The synthesized speech was subject to preliminary evaluation of the speech output and diagnostic annotation evaluation with the use of an Automatic Close Copy Speech (ACCS) synthesis tool as an audio screening procedure, the BOSS synthesis system for Polish was assessed in five speech quality judgment tests based on SAM/EAGLES standards.

After a continuous diagnostic evaluation during the development phase, the performance of the BOSS TTS system for Polish was formally assessed in a series of perception experiment comprising five speech quality judgment tests [22]. The tasks were either preference tests comparing two synthesized versions of the same sentence or assessments of naturalness and intelligibility expressed as mean opinion scores (MOS). MOS test stimuli were semantically unpredictable sentences (SUS), i.e. sentences that are grammatically correct but containing meaningless word sequences. The tests were administered to 20 Polish students of different philologies, 11 females and 9 males. The test sessions lasted approximately 60 minutes.

Two properties of the Polish speech corpus described in Sec. 3 were specifically targeted by the perception tests as factors affecting the synthetic speech quality: firstly, the inclusion of a large set of phrases covering CVC triphones in sonorant context and various intonation patterns (Base D); secondly, the inclusion of phrases covering the most frequent consonant structures.

The evaluation tests confirmed the synthetic speech quality is enhanced when the automatic annotation is manually corrected by trained phoneticians. However, the improvement was smaller than expected, indicating that in general the automatic annotation tools performed well. Currently software is being developed which will enhance the automatic segmentation engine and include automatic annotation of intonation structures too.

Including phonetically rich phrases in the unit selection corpus definitely boosts the quality of the synthetic speech. But it is worth noting that the inclusion of linguistic material designed to cover specific structures of the language also increases the corpus size. It has been reported in the literature that larger corpora produce better synthesis quality, everything else being equal. The effects of corpus size and structure are therefore likely to be confounded, and their relative contribution to perceptual quality reduction is difficult to quantify at this stage. Still, a diagnostic analysis of the unit selection procedure revealed that indeed the system preferably selects candidate units originating within consonant clusters when such units are available, instead of assembling clusters from individual consonants taken from other contexts. This is of course the desired behavior because, in natural speech, consonants in clusters exhibit stronger co-articulatory effects than consonants in simpler syllable structures, and good coverage of such effects by judicious corpus design is expected to pay off in terms of perceived higher synthesis quality.

5. Conclusions

This paper has presented work in progress on Polish text-to-speech synthesis. The BOSS speech synthesis system originally developed for German has served as a TTS architecture for implementing synthesis components required for the Polish language. A new predictive model for assigning an appropriate temporal structure to synthetic utterances was developed and demonstrated to yield very good performance. Moreover, cost functions for acoustic unit selection from the corpus to match a given target sentence specification were investigated and implemented.

For both new Polish synthesis components, the methodology applied is largely language-independent but the data used to train the duration model and the cost functions are necessarily language-dependent. The majority of the factors known to affect the duration of speech sounds are shared across languages; however, the relative impact of these factors and the details of their phonetic implementation are specific to the target language. In a similar way, a well-established set of phonetic features and acoustic dimensions is used in the cost functions to evaluate candidates during the acoustic unit selection process, whereas the language-specific details are implicit in the acoustic structure of the speech corpus data. So far, the extensible, multilingual design of the BOSS system architecture has been shown to be capable of accommodating modules for different languages. It needs to be tested further on components that implement structural differences between languages more directly than segmental duration and acoustic

unit selection, for instance modules for intonation modeling or morphological analysis.

Criteria for defining the structure of the speech corpus interact with unit selection criteria. Therefore, a large-scale evaluation is required to establish the optimal combination of TTS modules and unit selection algorithms. An international initiative, the Blizzard Challenge, aims to compare research techniques for corpus-based synthesis using the same corpus data [11]. As of now, the Polish version of BOSS has not participated in the Blizzard tasks, which have so far only addressed the English language. The parallel initiative of the European Center of Excellence for Speech Synthesis [10] attempts to evaluate not only entire TTS systems but also specific system components, and our research groups are actively involved in this initiative.

Future work concerning the technical components of the synthesis system will seek to further refine the cost functions and unit concatenation methods. Among the linguistic and phonetic components, a more sophisticated prosody control module will have to be implemented to alleviate the domain-specificity of the prosodic structure. With respect to the annotation techniques, it is intended to create tools that enable full automatisation of both segmental and prosodic annotation of Polish speech data for the needs of corpus-based synthesis. Work is ongoing to develop annotation procedures for expressive speech. The corpus structure and coverage will be elaborated in two respects: First, for neutral speech synthesis, additional linguistic material is intended to enhance coverage of co-articulatory effects in syntactically and phonetically rich sentences. Second, the corpus will be expanded for to cover expressive speech as well.

Acknowledgements. This research was supported by the Polish Ministry of Scientific Research and Information Technology, project no. R00 035 02. It has also been supported by an Alexander von Humboldt Polish Honorary Research Fellowship awarded to one of the authors (B.M.).

REFERENCES

[1] A.J. Hunt and A.W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database", *Proc. IEEE Int. Conf. on Acoustics and Speech Signal Processing* 1, 373–376 (1996).

[2] A.W Black and P. Taylor, "Automatically clustering similar units for unit selection in speech synthesis", *Proc. European Conf. on Speech Communication and Technology* 2, 601–604 (1997).

[3] A.P. Breen and P. Jackson, "Non-uniform unit selection and the similarity metric within BT's Laureate TTS system", *Proc. Third Int. Workshop on Speech Synthesis* 1, 373–376 (1998).

[4] M. Beutnagel, M. Mohri, and M. Riley, "Rapid unit selection from a large speech corpus for concatenative speech synthesis", *Proc. Eur. Con. on Speech Communication and Technology* 2, 607–610 (1999).

[5] A. Conkie, "Robust unit selection system for speech synthesis", *Collected Papers of the 137th Meeting of the Acoustical Society of America and the 2nd Convention of the European Acoustics Association: Forum Acusticum Berlin* 1, 1PSCB\10 (1999).

[6] M. Balestri, A. Pacchiotti, S. Quazza, P.L. Salza, and S. Sandri, "Choose the best to modify the least: a new generation concatenative synthesis system", *Proc. Eur. Conf. Speech Communication and Technology* 5, 2291–2294 (1999).

[7] N. Iwahashi, and Y. Sagisaka, "Speech segment network approach for an optimal synthesis unit set", *Computer Speech and Language* 9, 335–352 (1995).

[8] B. Möbius, "Rare events and closed domains: two delicate concepts in speech synthesis", *Int. J. Speech Technology* 6 (1), 57–71 (2003).

[9] J.P.H. Santen and A.L. Buchsbaum, "Methods for optimal text selection", *Proc. Eur. Conf. on Speech Communication and Technology* 2, 553–556 (1997).

[10] *ECESS: European Center of Excellence on Speech Synthesis*, http://www.ecess.eu (2008).

[11] *SYNSIG: Speech Synthesis Special Interest Group of ISCA*, http://www.synsig.org/index.php/Blizzard_Challengeil (2008).

[12] *BOSS: The Bonn Open Synthesis System*, http://www.ifk.uni-bonn.de/search?SearchableText=boss (2008).

[13] S. Breuer, "Multifunktionale und multilinguale Unit-Selection-Sprachsynthese – Designprinzipien fur Architektur und Sprachbausteine" , *Phd Thesis*, Universitat Bonn, Bonn, 2008.

[14] E. Klabbers and K. Stober, R. Veldhuis, P. Wagner, and S. Breuer, "Speech synthesis development made easy", *The Bonn Open Synthesis System* 1, 521–524 (2001).

[15] K. Stöber, T. Portele, P. Wagner, and W. Hess, "Synthesis by word concatenation", *Proc. Eur. Conf. on Speech Communication and Technology* 2, 619–622 (1999).

[16] *SAMPA for Polish Homepage*, http://www.phon.ucl.ac.uk/home/sampa/polish.htm (2008).

[17] W. Jassem, "Illustrations of the IPA", *Polish J. Int. Phonetic Association* 33 (1) 103–107 (2003).

[18] G. Demenko, M. Wypych, and E. Baranowska, "Implementation of grapheme-to-phoneme rules and extended SAMPA alphabet in Polish text-to-speech synthesis", *Speech and Language Technology* 7, 79–97 (2003).

[19] M. Szymański, and S. Grocholewski, "Semi-automatic segmentation of speech: manual segmentation strategy. Problem space analysis", *Advances in Soft Computing, Computer Recognition Systems* 1, 747–755 (2005).

[20] K. Sjölander and J. Beskow, *Wavesurfer*, http://www.speech.kth.se/wavesurfer/ (2008).

[21] K. Klessa, , M. Szymański, S. Breuer, and G. Demenko, "Optimization of Polish segmental duration prediction with CART", *6th ISCA Workshop on Speech Synthesis (SSW-6) Proc.* 1, CD-ROM (2007).

[22] G. Demenko, J. Bachan, B. Möbius, K. Klessa, M. Szymański, and S. Grocholewski, "Development and evaluation of Polish speech corpus for unit selection speech synthesis systems", *Proc.: Interspeech 2008* 1, CD-ROM (2008).

[23] S. Breuer, K. Francuzik, G. Demenko, and M. Szymański, "Analysis of Polish segmental duration with CART", *Proc. Speech Prosody Conf.* 1, CD-ROM (2006).

[24] S. Breuer and J. Abresch, Unit selection speech synthesis for a directory enquiries service", *Proc. ICPhS Barcelona 2003* 1, CD-ROM (2003).

[25] D. Gibbon and J. Bachan, "An automatic close copy speech synthesis tool for large-scale speech corpus evaluation", *Proc. Sixth International Language Resources and Evaluation (LREC'08)* 1, CD-ROM (2008).

[26] *ELDA: Evaluations and Language resources Distribution Agency*, http://www.elda.org/ (2008).