

The design of Polish Speech Corpus for Unit Selection Speech Synthesis

Grazyna Demenko,* Bernd Möbius,** and Katarzyna Klessa*

*Institute of Linguistics, Department of Phonetics
Adam Mickiewicz University, Poznań

**Institute of Natural Language Processing, University of Stuttgart,
and Institute of Communication Sciences, University of Bonn

lin@amu.edu.pl

bernd.moebius@ims.uni-stuttgart.de

klessa@amu.edu.pl

ABSTRACT

The Bonn Open Synthesis System (BOSS) is open-source software for 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 next section focuses on the implementation of Polish TTS modules in BOSS architecture (duration prediction and cost function) and the steps involved in preparing a new speech corpus for BOSS.

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.

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 a 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 stable 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 modes. 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 these speaking modes 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 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, and the implementation of Polish modules for duration prediction and cost functions for the BOSS system is discussed in section three 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:

1. 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.

2. Base B: All Polish diphones produced in 114 grammatically correct but semantically nonsense phrases.
3. Base C: Phrases with CVC triphones (in non-sonorant voiced context and with various intonation patterns). 676 phrases were recorded for triphone coverage.
4. 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 to 180 Hz).
5. 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.
6. 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 4 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).

With a view to simplifying the annotation of the pitch accents only two dimensions were considered (Figure 1): the pitch movement direction and its position with respect to accented syllable boundaries. 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).

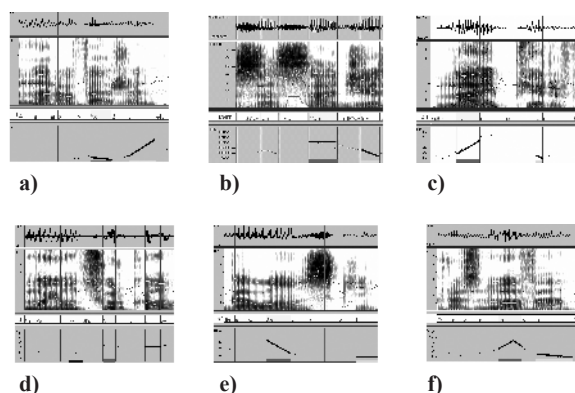


Figure 1. Pitch accents inventory: a) pitch movement with rising intonation R (on the post-accented syllable – LH); b) falling intonation F (on post-accented syllables HL); c) rising intonation on the accented syllable; d) level intonation; e) falling intonation on the accented syllable; f) rising – falling intonation on the accented syllable. Accented syllables are bolded.

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 were 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 adjoining 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-occurring 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 phonotactic 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. For the feature ‘Place of articulation’ 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 ‘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), pre-nuclear stress, no stress. The sound position within the phrase could be either initial, medial or final.

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 (Figures 2 and 3).

The utterance: *pociąg pośpieszny do Krotoszyna przez Malbork oraz Bydgoszcz wjedzie wyjątkowo na szósty peron na dworcu zachodnim* (Eng. *The fast train to*

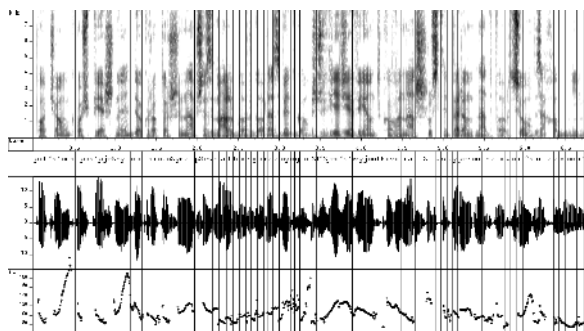


Figure 2. Spectrogram of the synthesized utterance:
pociąg pociąg pociąg do Krotoszyna przez peron na dworcu zachodnim.

Krotoszyn through Malbork and Bydgoszcz will be arriving today only at platform number six at the western station), was built from words, syllables, phonemes. The segmental features of this synthesized utterance were very good, the intonation was correct because the suprasegmental structure of database was representative enough.

Figure 3 shows the example of the synthesized utterance: *ta ruda panienska jest szwagierką Marylki*, (Eng. *that red-headed young lady is Marylka's sister-in-law*) with linguistic structures not contained in the database 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.

The synthesized speech was subject to preliminary evaluation of the speech output [22] and diagnostic annotation evaluation with the use of an Automatic Close Copy Speech (ACCS) synthesis tool [25] as an audio screening procedure, the BOSS synthesis system for Polish was assessed in five speech quality judgement tests based on SAM/EAGLES standards. The results of speech synthesis were very good for utterances containing triphones in various prosodic contexts. Relatively good results were obtained for intonation contour modeling.

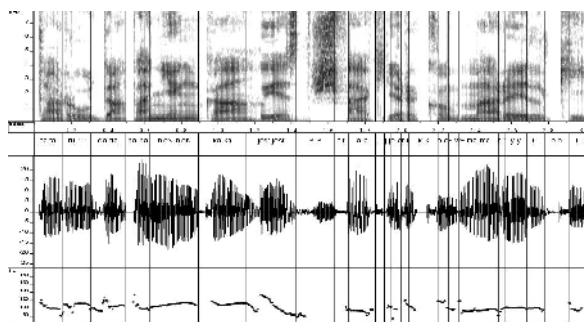


Figure 3. Spectrogram of the synthesized utterance:
ta ruda panienska jest szwagierką Marylki.

5. Discussion and Conclusion

As regards the technical solutions for the synthesis system it is planned to further develop the cost function and implement a more sophisticated prosody control module. Another necessary improvement is needed for the concatenation method and join costs. With respect to the annotation techniques, it is intended to create tools enabling full automation of both segmental and suprasegmental annotation of Polish speech data for the needs of unit selection. The work is going on developing tools for annotation of expressive speech. The database will be elaborated in two respects: first, for neutral speech synthesis improvement – new recordings of read speech will be provided using the text material covering approximately 10,000 Polish triphones in syntactically and phonetically rich sentences (prepared within the present project); second, the database will be expanded for expressive speech. Finally, it is planned to verify specifications of our speech corpus structure with ECESS guidelines and submit the database for validation and expertise by an external institution, for example ELDA [26].

The main disadvantage of corpus based synthesis is a lack of flexibility. Because the signal processing in it is either non-existent or limited, there is no possibility to change prosody and speaking style. Current speech databases allow for excellent re-synthesis in a fixed speaking style, but in spite of their size (10–30 hours) they are unable to produce different styles of speech. As comprehension is no longer an issue in speech synthesis nowadays, the most important questions seem to concern its naturalness and expressiveness. Speech synthesis results in BOSS system for Polish examples are available at: http://main.amu.edu.pl/~fonetyka/synthesis_examples.html.

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, Bernd Möbius.

BIBLIOGRAPHY

- [1] Hunt, A. J., Black, A. W., Unit selection in a concatenative speech synthesis system using a large speech database, Proceedings of the IEEE International Conference on Acoustics and Speech Signal Processing, Munchen, Germany, 1996, vol. 1, 373–376.
- [2] Black, A. W. and Taylor P., Automatically clustering similar units for unit selection in speech synthesis, Proceedings of the European Conference on Speech Communication and Technology (Rhodos, Greece), 1997, vol. 2, 601–604.
- [3] Breen, A. P. and Jackson, P., Non-uniform unit selection and the similarity metric within BT's Laureate TTS system, Proceedings of the Third International Workshop on Speech Synthesis, Jenolan Caves, Australia, 1998, 373–376.
- [4] Beutnagel, Mark and Mohri, Mehryar and Riley, Michael, Rapid unit selection from a large speech corpus for concatenative speech synthesis, Proceedings of the European Conference on Speech Communication and Technology, Budapest, Hungary, 1999, vol. 2, 607–610.
- [5] Conkie, A., 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, Germany, 1999, Paper 1PSCB/10.

- [6] Balestri, Marcello and Pacchiotti, Alberto and Quazza, Silvia and Salza, Pier Luigi and Sandri Stefano, Choose the best to modify the least: a new generation concatenative synthesis system, Proceedings of the European Conference on Speech Communication and Technology, Budapest, Hungary, 1999, vol. 5, pp. 2291–2294.
- [7] Iwahashi, Naoto and Sagisaka, Yoshinori, Speech segment network approach for an optimal synthesis unit set, *Computer Speech and Language*, 1995, vol. 9, pp 335–352.
- [8] Möbius, Bernd, Rare events and closed domains: Two delicate concepts in speech synthesis *International Journal of Speech Technology* 2003, vol. 6, n. 1, 57–71.
- [9] Santen, J.P.-H. and Buchsbaum, Adam~L., Methods for optimal text selection, Proceedings of the European Conference on Speech Communication and Technology (Rhodos, Greece), 1997, vol. 2, 553–556.
- [10] ECESS: European Center of Excellence on Speech Synthesis Web Page. Online: <http://www.ecess.eu/>, accessed on 15 December 2008.
- [11] SYNSIG: Speech Synthesis Special Interest Group of ISCA. Online: http://www.synsig.org/index.php/Blizzard_Challengeil, accessed on 15 December 2008.
- [12] BOSS: The Bonn Open Synthesis System. Online: <http://www.ifk.uni-bonn.de/search?SearchableText=boss>, accessed on 15 December 2008.
- [13] Breuer, S., Multifunktionale und multilinguale Unit-Selection-Sprachsynthese – Designprinzipien für Architektur und Sprachbausteine, PhdThesis Universität Bonn, 2008.
- [14] Klabbers, Esther and Stober, Karlheinz and Veldhuis, Raymond and Wagner, Petra and Breuer, Stefan, Speech synthesis development made easy: The Bonn Open Synthesis System, Proceedings of the European Conference on Speech Communication and Technology (Aalborg, Denmark), 2001, vol. 1, 521–524.
- [15] Stöber, K., Portele, T., Wagner, P., Hess, W., Synthesis by word concatenation, Proceedings of the European Conference on Speech Communication and Technology (Budapest, Hungary), 999, vol. 2, 619–622.
- [16] SAMPA for Polish – homepage, accessed on 15 December 2008: <http://www.phon.ucl.ac.uk/home/sampa/polish.htm>.
- [17] Jassem, W., Illustrations of the IPA. *Polish. Journal of the International Phonetic Association* vol. 33 (1) 103–107 2003.
- [18] Demenko, G., Wypych, M., and Baranowska, E., “Implementation of Grapheme-to-Phoneme Rules and Extended SAMPA Alphabet in Polish Text-to-Speech Synthesis”, *Speech and Language Technology*, vol. 7. [Ed.] PTFon, 79–97, Poznań, 2003.
- [19] Szymański, M. and Grocholewski, S., Semi-Automatic Segmentation of Speech: Manual Segmentation Strategy. *Problem Space Analysis, Advances in Soft Computing, Computer Recognition Systems*, 747–755, Springer Berlin, 2005.
- [20] Sjölander, K., Beskow, J. Wavesurfer homepage, accessed on 15 December 2008: <http://www.speech.kth.se/wavesurfer/>.
- [21] Klessa, K., Szymański, M., Breuer, S. and Demenko, G., Optimization of Polish Segmental Duration Prediction with CART”, 6th ISCA Workshop on Speech Synthesis (SSW-6) Proc., Bonn, 2007.
- [22] Demenko, G., Bachan J., Möbius B., Klessa K., Szymański M., Grocholewski S., Development and Evaluation of Polish Speech Corpus for Unit Selection Speech Synthesis Systems. Proceedings: Interspeech 2008, September 22–26, 2008, Brisbane, Australia.
- [23] Breuer, S., Francuzik (Klessa), K., Demenko, G., Szymański, M., Analysis of Polish Segmental Duration with CART, Proceedings of Speech Prosody Conference, Dresden. 2006.
- [24] Breuer, S., Abresch, J. 2003. Unit Selection Speech Synthesis for a Directory Enquiries Service. In Proceedings of the (ICPhS). Barcelona.

- [25] Gibbon, D., Bachan, J. 2008. An automatic close copy speech synthesis tool for large-scale speech corpus evaluation. In: Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Ed. European Language Resources Association (ELRA), 28–30 May 2008 Marrakech, Morocco.
- [26] ELDA: Evaluations and Language resources Distribution Agency. Online: <http://www.elda.org/>, accessed on 15 December 2008.