THE DESIGN OF POLISH SPEECH CORPORA FOR SPEECH SYNTHESIS IN BOSS SYSTEM

Grażyna Demenko[1] Katarzyna Klessa[1] Marcin Szymański[2] Jolanta Bachan[1]

[1] Institute of Linguistics, Adam Mickiewicz University, ul. Międzychodzka 5, 60-371Poznań
[2] Poznan University of Technology, Institute of Computing Science, ul. Piotrowo 2, 60-965 Poznań
e-mail: lin@amu.edu.pl, katarzyna.klessa@amu.edu.pl, mszymanski@cs.put.poznan.pl, jolabachan@gmail.com

Abstract. The present paper presents ongoing research and development of Polish unit selection speech synthesis based on a four-hour speech corpus. In the first section, the origins and workings of the unit selection method for speech synthesis are explained. The second section focuses on the speech corpus development and annotation. Section three details the implementation of Polish modules in BOSS synthesis system, namely cost function development and Polish phone duration prediction. In section four the results of the preliminary assessment of the quality of synthesized speech obtained with the Polish version of the BOSS Synthesizer are reported.

1. Introduction

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 resctricted 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. This means that these speaking modes should be quite uniform over the domains chosen. According to ECESS specification for expressive speech synthesis the dominant speaking mode should be that of the 'parliamentary speeches' domain and speech corpora should be composed as following: transcribed speech novels with short sentences, selected phrases, frequent phrases, triphone coverage sentences. The specification of subcorpora structure is highly language dependent and needs to be carefully prepared. Our ongoing research and development of Polish unit selection speech synthesis is based preliminarily only on a 4-hour corpus. Since among many factors affecting the quality of speech recognition and synthesis, prediction and modeling of prosody plays an essential role, we have paid a special attention to the analyses of CVC triphones in various prosodic contexts.

Therefore, the aims of the current research were: a) to establish an inventory of distinctive (with respect to realization and perception) intonation events to be annotated in a speech database for the application in speech synthesis, b) to analyse speech subcorpora structure for Polish for further specification and development, c) to implement Polish modules in BOSS.

1.1 BOSS

Bonn Open Synthesis System (BOSS) (Breuer, et al., 2000) is a general open-source framework for nonuniform unit-selection-based synthesis. The data is stored in two databases: the speech inventory and the SQL tables which contain information about the speech data. It is written in C ++ under the Linux platform.

BOSS architecture follows the client-server paradigm. The client is a program that can run remotely (on any system), sending input in the XML format to the server and receiving the speech signal. The server consists of independent modules, each performing a different synthesis step (e.g. duration prediction). This makes a

separate development and testing of the modules possible.

The unit selection speech synthesis systems like BOSS perform the synthesis using a two-stage algorithm: The first stage is the candidate selection, where the requested word units are selected from the corpus. Optionally, this step can be performed on the basis of some criteria. If a word is not available in the corpus, units can be searched at lower levels (there are two more levels available, syllable and the phoneme level).

In the second stage, a final unit selection is made. All possible candidates available in the corpus are used to form a graph, where the nodes represent candidate units, while the possible transitions between units (which correspond to adjacent sentence fragments) are represented by the arcs in the graph. Cost functions can be applied to both nodes and the edges or to the edges only. The unit selection dynamic-programming algorithm finds the best path through the graph by computing the minimal sum of cost values on a path between start and end node (Campbell, 1997).

Some systems additionally perform signal manipulation as the final stage. If they do so, it is usually needed at points where there is a fundamental frequency mismatch between concatenated units or when the required prosody could not be found in the database.

2. Polish speech corpus development

2.1 *Structure*

The entire speech material (4 hours) was read by a professional radio speaker during several recording sessions, supervised by an expert phonetician.

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 realized 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 to180Hz).
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: Contains 15 minutes of prose and newspaper articles.

2.2 *Segmental annotation*

The computer coding conventions were drawn up in SAMPA for Polish created by J. C. Wells (1996) with revisions and extensions and in the IPA alphabet (IPA Homepage, Jassem, 2003). 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. The input set of symbols for Polish was defined here as a set of the following symbols: X={a, ą, b, c, ć, d, e, ę, f, g, h, i, j, k, l, ł, m, n, ń, o, ó, p, q, r, s, ś, t, u, v, w, x, y, z, ź, ż, #, ##}. One hash substitutes inter-word spaces in a string of orthographic symbols, and two hashes denote sentence final punctuation marks. 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. Besides, 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 (Szymański, Grocholewski, 2005) has been developed for the automatic segmentation of speech. Its features include:

a) calculating segment (usually phoneme) boundaries based on phonetic transcription,
b) context-dependent phoneme duration models,
c) considering "forced" transition points for semi–automatic segmentation,
d) accepting triphone statistical models trained with HTK tools,

e) tools for duration models calculation,
f) orthographic-to-phonetic conversion,
g) evaluation of decision trees to synthesis unseen triphones,
h) accepting wave or MFCC files (plus several label formats) as input,
i) posterior triphone-to-monophone conversion.

2.3 Suprasegmental annotation

The annotation system for unit selection requires information about segmental and suprasegmental structure, such as phone, syllable and word boundaries, syllable stress, phrase boundaries of different type and strength.

For prosody modeling, only the fundamental types of prosodic structures were distinguished, such as word stress and phrase accent placement, accent type or phrase boundary type according to the BOSS label format - BLF (cf. Breuer et al. 2000).

After the automatic labeling with SALIAN software our database was manually verified and annotated for suprasegmental features by 4 experts on the basis of perceptual and acoustic analyses of the speech signals. On the phrase level information about sentence and intonation type was provided. On the syllable level pitch accent types were marked. Pitch accents are determined by pitch variations occurring on the successive vowels/syllables and pitch relations between syllables. The annotation of pitch accent types can be complex because it may include combinations of many acoustic features (e.g. pitch movement direction, range of the change of pitch, pitch peak position).

With a view to simplifying the annotation of the pitch accents we took into consideration only two features: direction of the pitch movement and its position with respect to accented syllable boundaries. The resulting inventory of pitch accent labels includes: 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).

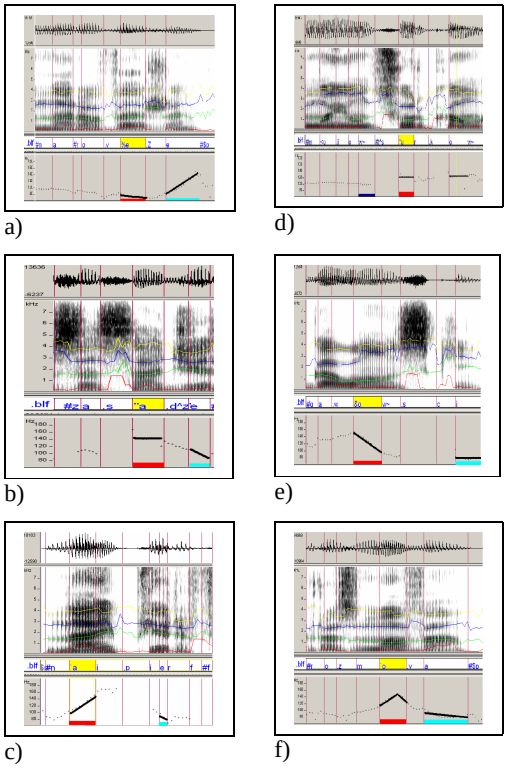

Fig. 1: The inventory of pitch accents: a) pitch movement with rising intonation R (on the post-accented syllable - LH) b) falling intonation F (on post-accented syllables HL), c) rising intonation on the accented syllable d) level intonation e) falling intonation on the accented syllable f) rising – falling intonation on the accented syllable. Accented syllables are marked in colour.

The rules of syllabifying in our research were based on the assumption that there is a relationship between *sonority* and *syllable structure*.

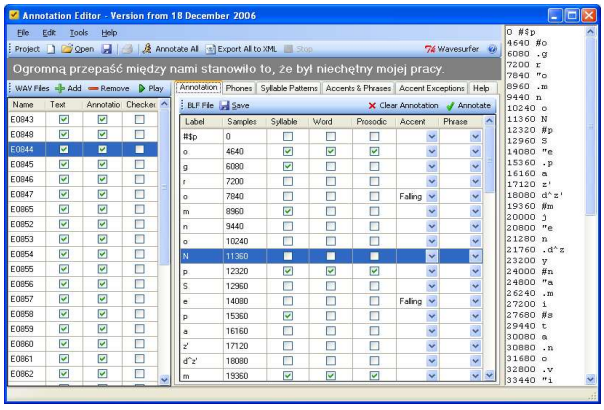

Fig. 2: Annotation Editor window.

For establishing syllable boundaries for Polish, the rules based on 20-million word Polish lexicon was set by an expert and then fully automatically implemented in software program: *Annotation*

Editor – a tool integrating the software for automatic division into syllables with segmentation programme, stress and accent control unit, and an editor of BLF files, and Wavesurfer (Sjölander, Beskow, 2006).

3. Implementation of Polish TTS modules in BOSS

So far, two proprietary Polish units were implemented for BOSS: the duration prediction module (see 3.1 below) 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, node cost function consists of the following components:

1. the absolute difference between the CART-predicted segment duration and the candidate unit duration (in ms),
2. the boolean difference between predicted and actual stress value, multiplied by 10,
3. the discrepancy regarding phrase type (question or indication, raising or falling intonation) and phrase location within a sentence (final or comma-terminated), multiplied by 20.

In the most recent implementation, the following features are considered by the transition cost function:

1. the Euclidean MFCC distance between the left segment right edge and the right segment left edge,
2. 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 penalized. (See *Future Works* for more remarks.)

3.1 Duration prediction

The first step in duration modeling for the purpose of the BOSS project was to provide a list of features potentially modifying the speech timing in Polish. The first version of the list was formulated on the basis of the subject literature. Polish segmental duration has been analysed with a view to speech technology applications some time ago (eg. Imiołczyk, et al., 1994), however most results were delivered using limited amount of speech data that mostly comprised of isolated word (or nonsense word) lists.

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 (cf. also Breuer et.al., 2006). 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:

1. 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).
2. 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.
3. The position within a higher unit of speech organization structure. (syllable, word, phrase).
4. Information about the direct neighborhood of the phone in question (within and across word boundary, relative to properties of adjacent sounds or sound clusters).
5. Word length and foot length.
6. Syllable length, phrase length, and the length of the whole source utterance.
7. 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 best results are shown in Table 1. (cf. also Klessa et al., 2007). They are quite good as compared to the results reported for other languages (eg. Chung, et al., 2001 for Korean, Batusek, 2002, for Czech).

Tab. 1: CART results for 57 feature vectors.

Percentage of heldout data for testing	*RMSE [ms] (Root Mean Squared Error)*	*Mean Correlation*	*Mean Error (abs)*
5%	15.4010	0.8080	11.3451 (10.4154)
10%	15.5124	0.8048	11.4217 (10.4966)
15%	15.6353	0.8013	11.5038 (10.5889)
20%	15.7082	0.7993	11.5724 (10.6221)

4. Evaluation

The BOSS synthesis system for Polish was assessed in two speech quality judgement tests. The preliminary assessment, Phase 1, was performed by Polish students of linguistics. In Phase 2, the speech output was assessed by a group of naive native speakers of Polish.

For Phase 1. fifteen sentences were used. Five of the sentences were re-synthesized original utterances from the corpus. The other ten synthesized utterances were new sentences containing common Polish vocabulary. The subjects' task was to assess intelligibility, naturalness and pleasantness of listening on a standard Mean Opinion Score (MOS) 5-point rating scale; 1 – indicating the lowest perceived quality, 5 – being the highest. The Phase 1 tests were carried out on eighteen 21-24 year-old Polish students of linguistics between. The students were divided into two smaller groups which took the tests separately in a quiet classroom. The stimuli were played twice via medium quality computer loudspeakers. After each presentation, the students had a pause of a few seconds to write down their scores. The test results are presented in Table 2.

The results show that the re-synthesized original sentences were assessed as being much better than synthetic speech. In all cases however, the students were rather hesitant in assigning the highest score. All the assessments of the re-synthesized original speech items received approximately one point more in the assessment than the synthesized items. Intelligibility was rated better than naturalness and pleasantness for both original and synthesized sentences.

Tab. 2: Results of the Phase 1 speech quality tests. I - intelligibility, N - naturalness, P - Pleasantness, re-SO - re-synthesized speech, S - synthesized. MOS score/5 - Mean Opinion Score out of 5, STDV - the standard deviation, Max and Min - maximal and minimal scores given by subjects.

	I		*N*		*P*	
	re-SO	*S*	*re-SO*	*S*	*re-SO*	*S*
MOS score/5	4.18	3.26	3.93	2.59	3.81	2.71
STDV	1.03	1.35	0.9	1.02	0.88	1.05
Max	5	5	5	5	5	5
Min	1	1	2	1	2	1

Phase 2 used 62 utterances, some of which were composed of more than one sentence. The testing procedure was the same as for Phase 1. The tests examined the synthesis of four parameters: consonant clusters, transitions between different phones and /j/ or /l/, phonetic contrasts and intonation.

The tests were taken by twelve people, six men and six women aged 9 to 61. The subjects took the tests separately or in pairs in a quiet room. The stimuli were played twice via medium quality laptop loudspeakers. (See results in Table 3.).

Tab. 3: Results for the Phase 2 speech quality tests. M - males, F - females.

	I		*N*		*P*	
	M	*F*	*M*	*F*	*M*	*F*
MOS score/5	3.59	3.81	3.31	3.37	3.60	3.20
		3.70		3.34		3.40
STDV	1.40	1.41	1.31	1.29	1.36	1.39
		1.40		1.30		1.37
Max	5	5	5	5	5	5
Min	1	1	1	1	1	1

Although none of the three sets received excellent scores, the scores over 3 points are encouraging for this stage of the BOSS system development for Polish. Intelligibility was graded best, but naturalness and pleasantness received only slightly worse scores overall.

The analyses of the synthetic speech samples show that the results of speech synthesis were very good for utternances containing triphones in various prosodic contexts. Relatively good results were obtained for intonation contour modeling. Figure 3. presents an example of a synthesized

utterance with a sophisticated melody using triphones from Base B (see p. 2.1. above).

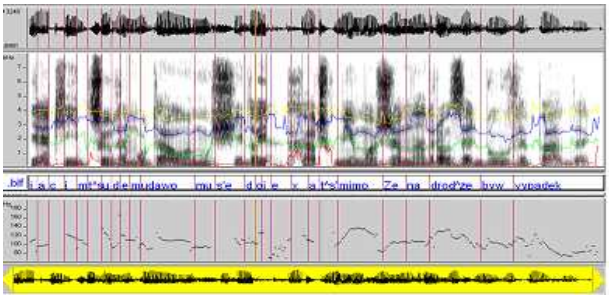

Fig. 3: Speech synthesis sample from Base B. The utterance: *jak mu sie udalo dojechać, mimo że na drodze był wypadek.*

5. Future works

Several features should be additionally taken into account for duration prediction as regards cost functions. Those include: syllable position within a phrase and unit left and right context. Moreover, splitting corpus sonorant-obstruent pairs should be avoided. Additional perception tests followed by careful verification as well as revision of problematic issues are required to improve the final quality of synthetic speech. Further extension of the speech corpora is also considered.

Acknowledgements

This research has been carried out under grant R00 035 02 received from the Polish Ministry of Scientific Research and Information Technology.

References

J. Adell, A. Bonafonte: Towards phone segmentation for concatenation speech synthesis, *5th Speech Synthesis Workshop, Pittsburgh, 2004.*

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

S. Breuer, K. Stober, P. Wagner, J. Abresch: Dokumentation zum Bonn Open Synthesis System BOSS II, *Unveroffentliches Dokument, IKP, Bonn, 2000.* http://www.ikp.uni-bonn.de/dt/forsch/phonetik/boss/BOSS_Documentation.pdf

S. Breuer, K. Francuzik, M. Szymański, G. Demenko, Analysis of Polish Segmental Duration with CART, *Proceedings of Speech Prosody Conference 2006, paper ID: 264, Dresden, 2006.*

N. Campbell, A. Black: Prosody and the selection of source units for concatenative synthesis. *In J. van Santen, R. Sproat, J. Olive, and J. Hirschberg (Eds.), Progress in Speech Synthesis, New York, Springer Verlag, (1997), 279-292.*

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

G. Demenko, A. Wagner: Analysis of accented syllables in different prosodic contexts for use in *Unit Selection Speech Synthesis, Archives of Acoustics, vol.30, 3. 2006.*

G. Demenko, M. Wypych, E. Baranowska, Implementation of Grapheme-to-Phoneme Rules and Extended SAMPA Alphabet in Polish Text-to-Speech Synthesis, *Speech and Language Technology, ed. PTFon, 2003, vol.7, pp. 79-97, Poznań.*

ECESS (European Center of Excellence on Speech Synthesis) Homepage: http://www.ecess.eu/

J. Imiołczyk, I. Nowak, G. Demenko: High intelligibility text-to-speech synthesis for Polish. *Archieves of Acoustics vol. 19 (2) str. 161-172, 1994.*

IPA (The International Phonetic Association) Homepage, http://www.arts.gla.ac.uk/IPA/ipa.html

W. Jassem: Illustrations of the IPA. Polish. *Journal of the International Phonetic Association vol. 33 (1) 103-107 2003.*

E. Klabbers, K. Stoeber, R. Veldhuis, P. Wagner and S. Breuer: Speech synthesis development made easy: The Bonn open synthesis system. *Proceedings of Eurospeech, 2001.*

K. Klessa, M. Szymański, S. Breuer, G. Demenko: Optimization of Polish Segmental Duration. Prediction with CART. *Proceedings of the 6th ISCA Workshop on Speech Synthesis (SSW-6), Bonn,2007.*

P. Mertens: The Prosogram: Semi-Automatic Transcription of Prosody based on a Tonal Perception Model, *B. Bel & I. Marlien (eds.) in Proceedings of Speech Prosody 2004, Nara (Japan).*

D. H. Milone, A. J. Rubio: Prosodic and Accentual Information for Automatic Speech Recognition, *Proceedings of IEEE, vol.11., no.4, July 2003, pp. 321-333.*

S. Narayanan, A. Alwan: Text to Speech Synthesis, New Paradigms and advances, *IMSC Press Multimedia Series, New Jersey, 2004.*

Sjölander, K., Beskow, J., 2006. Wavesurfer Homepage: http://www.speech.kth.se/wavesurfer/

M. Szymański, S. Grocholewski, Semi-automatic segmentation of speech: manual segmentation strategy; problem space analysis, A*dvances in Soft Computing, Computer Recognition Systems, Springer Verlag, 2005, pp.747-755.*

TTS examples for the present project: http://www.ppnt.poznan.pl/ltjim/