Alp-Synth: Unit Selection Slovenian Text-to-Speech Synthesis

Jerneja Žganec Gros, Aleš Mihelič, Mario Žganec

Alpineon razvoj in raziskave, d.o.o.
info@alpineon.com
http://www.alpineon.com

Abstract - The contribution focuses on the description of AlpSynth, a novel unit-selection driven Slovenian text-to-speech synthesis (TTS) system that is being built at Alpineon. We describe the design procedures of the TTS system and compare the methods implemented in AlpSynth to those applied within other Slovenian TTS systems.

A vital part of speech technology applications in modern voice application platforms is a text-to-speech engine. Text-to-speech synthesis enables automatic conversion into spoken form of any available textual information. The large number of successfully deployed speech technology applications has proven the technology works, cost savings or increased revenues have been achieved and speech recognition has introduced substantial improvements in the user interface over the touch-tone pad [1].

The initial attempts towards Slovenian TTS were mainly based on concatenation of diphones, and they resulted in a few demonstration systems [2], [3], [4], [5] and some first carrier-grade voice applications [6]. In [7] the authors describe the formation of only a text corpus for Slovenian corpus based TTS.

The AlpSynth TTS system follows principles similar to the S5 TTS system [3] and the Phonectic TTS system [6]. As the Phonectic TTS system, it is based on concatenation of basic speech units, extracted from a large speech corpus, instead of diphones only. Alp-Synth TTS has been completely rewritten towards achieving a minimum memory footprint and is based on publicly available information. It operates on a newly designed speech corpus that pays special attention to diphtong sequences and uses a novel approach to unit selection.

The input text is transformed into its spoken equivalent by a series of modules. A grapheme-to-allophone module produces strings of phonetic symbols based on information in the written text. The problems it addresses are thus typically language-dependent. A prosodic generator assigns pitch and duration values to individual phones. Pitch modeling is based primarily on predicting the proper Slovenian tonemic accent. Phone duration is predicted by a two level approach, taking into account how acceleration or slowing down affect the duration of individual phones. Final speech synthesis is based on unit selection and final concatenation. Given an input sequence of phonetic symbols a rather sophisticated unit selection algorithm that we are designing first selects the segments to be concatenated. It takes into account a number of criteria ranging from more and less preferred allophones for concatenation, the length and phonetic contexts of polyphones, spectral discontinuities, etc.

For unit selection text-to-speech synthesis a speech corpus of recorded and annotated elemental speech units is required [9], [10]. The quality of the output synthetic speech depends crucially on the quality of the speech corpus [8]. The longer elemental speech units are used the better and more natural-sounding synthetic speech the TTS system can yield. However, with longer elemental speech units the corpus size increases dramatically. Therefore, a compromise between the size of the speech corpus and the quality of the resulting speech has to be taken [12].

First an extended analysis of the frequency of Slovenian polyphone sequences was performed, based on the one described in [13], further expanded with the analysis of diphtong sequences. Large Slovenian text corpora have been transcribed into allophone sequences and statistically processed. The texts for the spoken corpus were selected by an optimization process optimizing the number of the most frequent polyphones covered by the spoken text and a minimum amount of the text to be read by the speaker. In order to be better able to compare the efficiency of the Phonectic TTS and the Alp-Synth TTS system, the same reader as for the Phonectic TTS system is chosen to pronounce the new expanded Alp-Synth speech corpus.

ACKNOWLEDGEMENT

The authors of the contribution wish to thank the Slovenian Ministry of Education, Science and Sports, the Slovenian Ministry of Information Society and the Slovenian Defense Ministry for co-funding the work carried out within the projects (CRP-2003 project No. V2-0896 and CRP-2004 project No. M2-0019).

REFERENCES

[1] W. Meisel, “Looking back at 2001 and forward to 2002”, *Speech recognition update*, p. 103, 2002.

[2] J. Gros, “Samodejno tvorjenje govora iz besedil”, *Linguistica et Philologica*, ZRC SAZU, Ljubljana, Slovenia, 2001. (in Slovenian)

[3] T. Šef, “Analiza besedila v postopku sinteze slovenskega govora”, *PhD Thesis*, Faculty of Computer Science and Informatics, University of Ljubljana, 2001. (in Slovenian)

[4] B. Vesnicer, N. Pavešić and F. Mihelič, “Korpusna sinteza govora”, *Proceedings of the ERK'01 Conference*, Vol. B, p. 253, Portorož, Slovenia, 2001. (in Slovenian)

[5] B. Vesnicer, “Umetno tvorjenje govora z uporabo prikritih Markovovih modelov”, *MSc Thesis*, Faculty of Electrical Engineering, University of Ljubljana, 2003. (in Slovenian)

[6] J. Gros, M. Žganec, A. Mihelič, M. Knez, A. Merčun, D. Marinčič, “The phonetic family of voice-enabled products”, *Proceedings of the conference Jezikovne tehnologije*, p. 127, Ljubljana, 2002.

[7] M. Rojc and Z. Kačič, “Design of optimal Slovenian speech corpus for use in the concatenative speech synthesis system”, *Proceedings of the Second international conference on language resources and evaluation*, p. 321, Athens, Greece, 2000.

[8] M. Beutnagel, M. Mohri, M. Riley, “Rapid unit selection from a large speech corpus for concatenative speech synthesis”, *Proceedings of the Eurospeech '99 Conference*, Budapest, Hungary, 1999.

[9] A. Conkie, “Robust unit selection system for speech synthesis”, *Procedings of the Eurospeech '99 Conference*, Budapest, Hungary, 1999.

[10] A. Conkie, M. Beutnagel, A. Syrdal and P. Brown, “Preselection of candidate units in a unit selection-based Text-to-Speech synthesis system”. *Proceedings of the ICSLP '00 Conference*, Beijing, China, 2000.

[11] M. Beutnagel, M. Mohri and M. Riley, “Rapid unit selection from a large speech corpus for concatenative speech synthesis”, *Proceedings of the Eurospeech '99 Conference*, Budapest, Hungary, 1999.

[12] J. Yi,“Natural-sounding speech synthesis using variable-length units”. *MEE Thesis*, Massachussetts Institute of Tehnology, 1998.

[13] A. Mihelič, “Zbirka govornih signalov za sintezo slovenskega govora”, *MSc Thesis*, Faculty of Electrical Engineering, University of Ljubljana, 2002. (in Slovenian)