Signal Processing in Text-to-Speech Synthesis

Advanced Signal Processing Seminar SS 2010

Harald Romsdorfer

Signal Processing and Speech Communication Laboratory
Graz University of Technology, Austria

Abstract

The last years showed a constant improve of the quality of text-to-speech synthesis systems. Applying text-to-speech synthesis is becoming more and more popular in commercial devices, like car nagivation systems, smart phones, or ebook readers. In this seminar, we want to give an overview of the current status of text-to-speech synthesis approaches with a focus on speech signal processing and statistical approaches.

1 Introduction

This is a LaTeX template for the text that you should prepare together with the presentation slides for the Advanced Signal Processing Seminar SS 2010. Please change author name, title and content according to your needs.

The following commands will generate a pdf file from your LaTeX text:

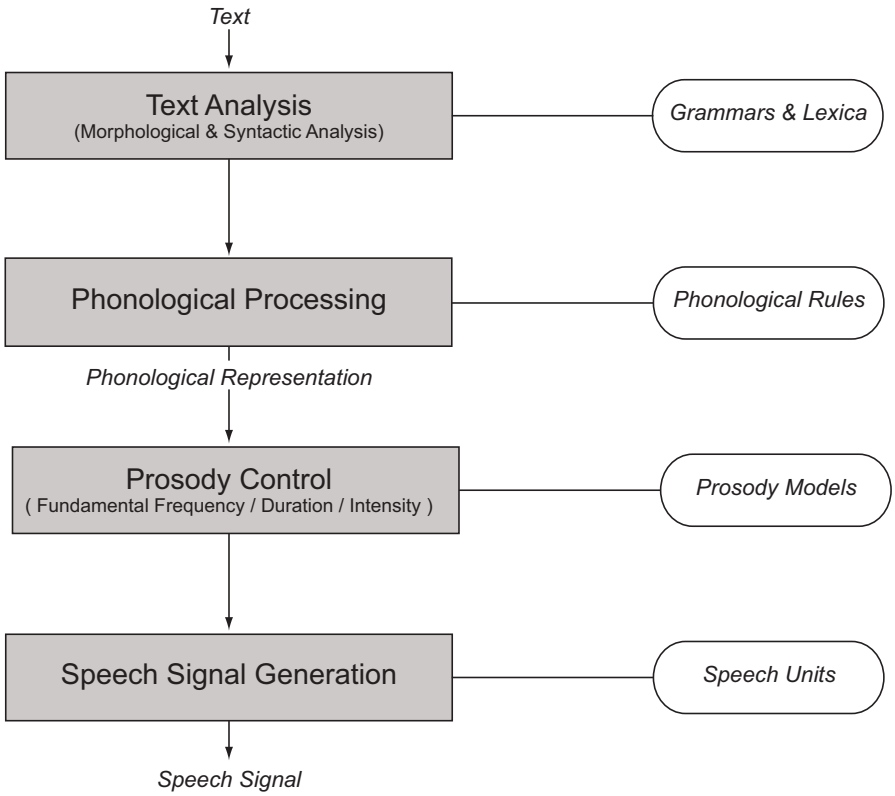
```
latex AdvSE2.tex
bibtex AdvSE2
latex AdvSE2.tex
latex AdvSE2.tex

dvips AdvSE2.dvi
ps2pdf AdvSE2.ps
```

A very good introduction to typesetting with LaTeX can be found in (1) and in the "classic" book of (2).

1.1 Administrative Information

- *I*nstructor: Harald Romsdorfer
- *M*eeting Date/Time: Fr 25.6., 16:00 - 19:00 and Di 29.6., 16:00 - 19:00 in meeting room IDEG134
- *M*ode: Each group of 1-2 students should select one topic. They should give an in-depth presentation of this topic and the referenced work therein (for about 1 - 1.5 hours). Following the presentation, we would like to discuss the presented topic. Therefore, it is necessary that each participant reads the presented article.
- *G*rading: Grades are given based on the presentation and the participation in discussions (50% :: 50%). The presentation slides and presentation text (approx. 6-10 pages) must be sent before the presentation to `romsdorfer@tugraz.at`.

1.2 Text-to-Speech Synthesis Overview



2 Topics

2.1 Analysis of Speech Signals (Feature Extraction)

- STFT, MFFC, RASTA, PLP, LSF
- Cepstral Smoothing
- Phone Clustering

see (3; 4)

2.2 Speech Corpus Segmentation

- HMM-based Forced Alginment
- DTW-based Pattern Matching

see (5; 6)

2.3 Prosodic Modification of Speech Signals

- Source-Filter Models: LPC, Spectral Coeffs (STRAIGHT)
- TD-PSOLA
- FD-PSOLA

see (7; 8)

2.4 Prosody Generation

- F0 Modeling
- Segment Duration Modeling
- Intensity Modeling

see (9; 10; 11)

2.5 Concatenative Speech Synthesis

- Diphone Synthesis
- Unit Selection Synthesis
- Source-Filter Synthesis: LPC, Fourier, MFCC

see (12; 13; 14; 15)

2.6 Statistical Parametric Speech Synthesis

- HMM-based Speech Synthesis: MFCC, LPC

see (16; 3)

2.7 Voice Transformation Algorithms

- Statistical Parametric Speech Synthesis
- Concatenative Speech Synthesis

see (17; 18; 19)

2.8 Language Transformation Algorithms

- Statistical Parametric Speech Synthesis
- Concatenative Speech Synthesis

see (20; 21)

2.9 Polyglot Speech Synthesis

- Foreign Inclusion Detection
- Polyglot Speech Synthesis: Concatenation-based, HMM-based

see (22; 23; 24; 11; 20)

3 Literature

Beside of the literature references given with each topic, the following books can be recommended for a more detailed overview of speech processing and especially of speech synthesis:

- Corpus-based speech synthesis (15)
- Speech and Language Processing (4)
- Text-to-Speech Synthesis (3)
- The HTK Book (5)

References

[1] T. Oetiker, H. Partl, I. Hyna, and E. Schlegl, *The Not So Short Introduction to LaTeX*, 3rd ed., August 2001.

[2] H. Kopka and P. W. Daly, *Guide to LaTeX, Fourth Edition*. Reading, MA: Addison-Wesley, 2004.

[3] P. Taylor, *Text-to-Speech Synthesis*, 2009.

[4] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, second edition ed. Pearson, 2009.

[5] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*. Cambridge: Cambridge University Engineering Departement, 2002.

[6] J.-P. Hosom, "Speaker-independent phoneme alignment using transition-dependent states," *Speech Communication*, 2008.

[7] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communications*, vol. 9, no. 5–6, pp. 453–467, December 1990.

[8] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," in *Proceedings of 2nd MAVEBA*, Firenze, Italy, September 2001, pp. 13–15.

[9] A. W. Black and A. Hunt, "Generating F_0 contours from the ToBI labels using linear regression," in *Proceedings of ICSLP'96*, 1996, pp. 1385–1388.

[10] J. Yamagishi, H. Kawai, and T. Kobayashi, "Phone duration modeling using gradient tree boosting," *Speech Communication*, vol. 50, no. 5, pp. 405–415, May 2008.

[11] H. Romsdorfer, "Polyglot text-to-speech synthesis. text analysis and prosody control," Ph.D. dissertation, No. 18210, Computer Engineering and Networks Laboratory, ETH Zurich (TIK-Schriftenreihe Nr. 101), January 2009.

[12] A. W. Black and N. Campbell, "Optimising selection of units from speech databases for concatenative synthesis," in *Proceedings of Eurospeech'95*, Madrid, Spain, September 1995, pp. 581–584.

[13] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proceedings of ICASSP'96*, Atlanta, Georgia, USA, 1996, pp. 373–376.

[14] J. P. H. van Santen, "Combinatorial issues in text-to-speech synthesis," in *Proceedings of Eurospeech'97*, Rhodes, Greece, September 1997, pp. 2511–2514.

[15] T. Dutoit, "Corpus-based speech synthesis," in *Handbook of Speech Processing*, J. Benesty, M. Sondhi, and Y. Huang, Eds. Berlin Heidelberg: Springer, 2008, ch. 23, pp. 471–487.

[16] J. Yamagishi, "An introduction to HMM-basd speech synthesis," Tokyo Institute of Technology, Tech. Rep., October 2006.

[17] M. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proceedings of ICASSP 1998*, 1998, pp. 285–288.

[18] ——, "Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction," in *Proceedings of ICASSP 2001*, Salt Lake City, Utah, USA, May 2001.

[19] J. Yamagishi, "Thousands of voices for HMM-based speech synthesis," in *IEEE Audio, Speech, & Language Processing*, 2009.

[20] J. Latorre, K. Iwano, and S. Furui, "Cross-language synthesis with a polyglot synthesizer," in *Proceedings of Interspeech 2005*, Lisbon, Portugal, September 2005, pp. 1477–1480.

[21] ——, "New approach to polyglot synthesis: How to speak any language with anyone's voice," in *ISCA Tutorial and Research Workshop on Multilingual Speech and Language Processing (MultiLing 2006)*, Stellenbosch, South Africa, April 2006.

[22] R. W. Sproat and J. T. Olive, "A modular architecture for multi-lingual text-to-speech synthesis," in *Proceedings of ESCA/IEEE Workshop on Speech Synthesis*, New Paltz, New York, September 1994, pp. 187–190.

[23] A. W. Black and K. A. Lenzo, "Multilingual text-to-speech synthesis," in *Proceedings of the ICASSP 2004*, Montreal, Canada, 2004.

[24] A. W. Black and T. Schultz, "Speaker clustering for multilingual synthesis," in *ISCA Tutorial and Research Workshop on Multilingual Speech and Language Processing (MultiLing 2006)*, Stellenbosch, South Africa, April 2006.