Improving Speech Synthesis – from Analysis to Implementation

Bettina Braun and Dominika Oliver

June 19, 2003

1 Overview

Our project will focus on possible improvements to standard speech synthesis systems. More specifically, we have the following research questions:

- How can we adapt speech synthesis systems to improve their output in different contexts and situations? This involves some knowledge about *natural* speech which is why we plan to analyse existing resources (partly using data-intensive techniques).
- What are the problems when evluating the changes to a baseline system?

Our project can be linked to most other existing projects and there is no need for any prior knowledge in phonetics, prosody, or speech synthesis. We mainly address students who are generally interested in speech synthesis.

2 Background

First, we present the state-of-the art in the field and discuss the difference between "rule-based" (diphone synthesis using MBROLA) and "statistical" (unit-selection) systems. Since unit selection systems don't allow modifications we shall focus on diphone synthesis, either MARY (http://mary.dfki.de/) or Festival [1].

Depending on the number of interested students we will build different discussion groups that will work on possible extensions to the existing baseline-system, e.g. various speech tempo or emotions (which might be interesting for the user-modelling group), information-structure (which might go together with both the generation and planning group), contrast and surprise (which are conditions in the clarification project.)

3 Analysis

Knowledge about the prominent phonetic features in speech can be gained by perception tests (with carefully controlled, often (re)synthesised stimuli) or by analysing *natural* speech. In this project, we will focus on the latter. We propose to analyse the influence of tempo, intonation patterns and emotional colouring on speech.

3.1 Tempo

In the last couple of years there has been increasing interest in duration models for speech synthesis, either rule-based or data-driven (e.g. [2]). There are many areas that are affected by changes in articulation rate: on the segmental level, phones may be reduced or deleted entirely [3] (which obviously has effects on the syllabic structure as well), on the suprasegmental level, accents may be truncated, merge or be realised with a limited range or duration.

For a corpus study, we propose the Kiel Corpus of Read Speech which is annotated for different speech rates [4]

3.2 Intonation

Influence of information structure on accentuation, alignment issues, internal structure of accents. The following databases (PoInt [5], Boston University Radio News Corpus [6], and data from the authors) will be investigated.

3.3 Emotion

We shall mainly focus on "surprise" for which data will be available from the "clarification questions"-project. An interesting overview is found in [7]

4 Implementation

In the practical part, we shall implement or suggest plausible solutions in an existing system (up to now MARY which has MBROLA-interface, see section 4.1). Depending on the interests of the students, we can discuss additions to existing tag-sets of mark-up languages (as this would be necessary as an interface to other projects that would like to have speech output).

4.1 MBROLA

The MBROLA input format needed for communicating with the synthesiser is as follows: Each line contains a phoneme name, a duration (in ms), and a series (possibly none) of pitch pattern points composed of two integer numbers each : the position of the pitch pattern point within the phoneme (in % of its total duration), and the pitch value (in Hz) at this position. Hence, the first line tells us that the phoneme /v/ starts with

phonemes	duration	pitch pattern
v	65	(0,189)
I	52	
l	46	(61,169)
k	88	
O	80	(50,215)
m	70	
@	56	
n	47	

Table 1: Example of MBROLA-format

a frequency of 189 Hz (reached at 0% of the phoneme). Pitch pattern points define a piecewise linear pitch curve.

5 Evaluation of Implementation

The evaluation is one crucial part of the project. We shall discuss how to build an evaluation scheme in order to assess the quality of our changes.

6 Preliminary Schedule

- Short introduction in Speech Synthesis Systems (MARY and Festival)
- Short literature review on tempo, intonation, emotion (depending on students' interests)
- Choice of corpus/corpora
- Data extraction (perl scripts, in the worst case own annotation)
- Build hypotheses for improvements
- Implementing findings using MBROLA

References

[1] R. Black and P. Taylor. (1998): Festival speech synthesis system: system documentation (1.1.1). Technical Report HCRC/TR-83. University of Edinburgh.

[2] Brinckmann, C. and J. Trouvain (2003): The role of duration models and symbolic representation for timing in synthetic speech. International Journal of Speech Technology 6, pp. 21–31.

[3] Trouvain, J., Koreman, J., Erriquez, A. and B. Braun (2001): Articulation rate measures and their relations to phone classification of spontaneous and read German speech. Proceedings ISCA Workshop on "Adaptation Methods for Speech Recognition", Sophia Antipolis (France), pp. 155-158.

[4] IPDS (1994). The Kiel Corpus of Read Speech. Vol I. Kiel CD#1. Retrieved April 10, 2002, from http://www-ipds/uni-kiel.de/publikationen/cd1.en.html.

[5] Karpinski, M. and J. Klesta. (2001): The project of an intonational database for the Polish language. Proc. Prosody 2000. pp. 113–118.

[6] M. Ostendorf, P. J. Price and S. Shattuck-Hufnagel: The Boston University Radio News Corpus, Boston University Technical Report No. ECS-95-001, March 1995.

[7] Schröder, M. (2001): Emotional Speech Synthesis - A Review. Proc. Eurospeech 2001, Aalborg, Vol. 1, pp. 561-564.