

Experiments with Synthesis of Swedish Dialects

Beskow, J. and Gustafson, J.

Department of Speech, Music & Hearing, School of Computer Science & Communication, KTH

Abstract

We describe ongoing work on synthesizing Swedish dialects with an HMM synthesizer. A prototype synthesizer has been trained on a large database for standard Swedish read by a professional male voice talent. We have selected a few untrained speakers from each of the following dialectal region: Norrland, Dala, Göta, Gotland and South of Sweden. The plan is to train a multi-dialect average voice, and then use 20-30 minutes of dialectal speech from one speaker to adapt either the standard Swedish voice or the average voice to the dialect of that speaker.

Introduction

In the last decade, most speech synthesizers have been based on prerecorded pieces of speech resulting in improved quality, but with lack of control in modifying prosodic patterns (Taylor, 2009). The research focus has been directed towards how to optimally search and combine speech units of different lengths.

In recent years HMM based synthesis has gained interest (Tokuda et al., 2000). In this solution the generation of the speech is based on a parametric representation, while the grapheme-to-phoneme conversion still relies on a large pronunciation dictionary. HMM synthesis has been successfully applied to a large number of languages, including Swedish (Lundgren, 2005).

Dialect Synthesis

In the SIMULEKT project (Bruce et al., 2007) one goal is to use speech synthesis to gain insight into prosodic variation in major regional varieties of Swedish. The aim of the present study is to attempt to model these Swedish varieties using HMM synthesis.

HMM synthesis is an entirely data-driven approach to speech synthesis and as such it gains all its knowledge about segmental, intonational and durational variation in speech from training on an annotated speech corpus. Given that the appropriate features are annotated and made available to the training process, it is

possible to synthesize speech with high quality, at both segmental and prosodic levels. Another important feature of HMM synthesis, that makes it an interesting choice in studying dialectal variation, is that it is possible to adapt a voice trained on a large data set (2-10 hours of speech) to a new speaker with only 15-30 minutes of transcribed speech (Watts et al., 2008). In this study we will use 20-30 minutes of dialectal speech for experiments on speaker adaptation of the initially trained HMM synthesis voice.

Data description

The data we use in this study are from the Norwegian Språkbanken. The large speech synthesis database from a professional speaker of standard Swedish was recorded as part of the NST (Nordisk Språkteknologi) synthesis development. It was recorded in stereo, with the voice signal in one channel, and signal from a laryngograph in the second channel.

The corpus contains about 5000 read sentences, which add up to about 11 hours of speech. The recordings manuscript was based on NST's corpus, and the selection was done to make them phonetically balanced and to ensure diphone coverage. The manuscripts are not prosodically balanced, but there are different types of sentences that ensure prosodic variation, e.g. statements, wh-questions, yes/no questions and enumerations.

The 11 hour speech database has been aligned on the phonetic and word levels using our Nalign software (Sjölander & Heldner, 2004) with the NST dictionary as pronunciation dictionary. This has more than 900.000 items that are phonetically transcribed with syllable boundaries marked. The text has been part-of-speech tagged using a TNT tagger trained on the SUC corpus (Megyesi, 2002).

From the NST database for training of speech recognition we selected a small number of unprofessional speakers from the following dialectal areas: Norrland, Dala, Göta, Gotland and South of Sweden. The data samples are considerably smaller than the speech synthesis database: they range from 22 to 60 minutes,

compared to the 11 hours from the professional speaker.

HMM Contextual Features

The typical HMM synthesis model (Tokuda et al., 2000) can be decomposed into a number of distinct layers:

- At the acoustic level, a parametric source-filter model (MLSA-vocoder) is responsible for signal generation.
- Context dependent HMMs, containing probability distributions for the parameters and their 1st and 2nd order derivatives, are used for generation of control parameter trajectories.
- In order to select context dependent HMMs, a decision tree is used, that uses input from a large feature set to cluster the HMM models.

In this work, we are using the standard model for acoustic and HMM level processing, and focus on adapting the feature set for the decision tree for the task of modeling dialectal variation.

The feature set typically used in HMM synthesis includes features on segment, syllable, word, phrase and utterance level. Segment level features include immediate context and position in syllable; syllable features include stress and position in word and phrase; word features include part-of-speech tag (content or function word), number of syllables, position in phrase etc., phrase features include phrase length in terms of syllables and words; utterance level includes length in syllables, words and phrases.

For our present experiments, we have also added a speaker level to the feature set, since we train a voice on multiple speakers. The only feature in this category at present is dialect group, which is one of *Norrland*, *Dala*, *Svea*, *Göta*, *Gotland* and *South of Sweden*.

In addition to this, we have chosen to add to the word level a morphological feature stating whether or not the word is a compound, since compound stress pattern often is a significant dialectal feature in Swedish (Bruce et al., 2007). At the syllable level we have added explicit information about lexical accent type (accent I, accent II or compound accent).

Training of HMM voices with these feature sets is currently in progress and results will be presented at the conference.

Acknowledgements

The work within the SIMULEKT project is funded by the Swedish Research Council 2007-2009. The data used in this study comes from Norsk Språkbank (<http://sprakbanken.uib.no>)

References

- Bruce, G., Schötz, S., & Granström, B. (2007). SIMULEKT – modelling Swedish regional intonation. *Proceedings of Fonetik, TMH-QPSR*, 50(1), 121-124.
- Lundgren, A. (2005). *HMM-baserad talsyntes*. Master's thesis, KTH, TMH, CTT.
- Megyesi, B. (2002). *Data-Driven Syntactic Analysis - Methods and Applications for Swedish*. Doctoral dissertation, KTH, Department of Speech, Music and Hearing, KTH, Stockholm.
- Sjölander, K., & Heldner, M. (2004). Word level precision of the NALIGN automatic segmentation algorithm. In *Proc of The XVIIth Swedish Phonetics Conference, Fonetik 2004* (pp. 116-119). Stockholm University.
- Taylor, P. (2009). *Text-To-Speech Synthesis*. Cambridge University Press.
- Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., & Kitamura, T. (2000). Speech parameter generation algorithms for hmm-based speech synthesis. In *Proceedings of CASSP 2000* (pp. 1315-1318).
- Watts, O., Yamagishi, J., Berkling, K., & King, S. (2008). HMM-Based Synthesis of Child Speech. *Proceedings of The 1st Workshop on Child, Computer and Interaction*.