HMM based speech synthesis system for Swedish Language

Bajibabu Bollepalli, Jonas Beskow, Joakim Gustafson

Department of Speech, Music and Hearing, KTH, Stockholm
{bajibabu, beskow}@kth.se, jocke@speech.kth.se

Abstract
This paper describes the Hidden Markov models-based Text-to-Speech system (HTS) for Swedish language. In this system, speech is represented by a set of parameters. This scheme of representation enables us to modify these parameters, and reconstruct speech from them. Due to this flexibility, it is possible to generate desired voice qualities, e.g., breathy voice, creaky voice, etc. Here, we describe the architecture of the current state-of-the-art HTS system and present preliminary evaluation results for Swedish language.

1. Introduction
The aim of a speech synthesis system is to generate a human like voice from arbitrary text. These systems have been under development for several decades. Recent progress in speech synthesis has produced synthesizers with high intelligibility and naturalness. Traditional concatenative or unit-selection based speech synthesis systems (A. Hunt and A. Black 1996), which synthesize the speech by joining different length speech-units (like phones, di-phones and syllables, etc.) derived from the natural speech, requires large amount of training data to synthesis good quality of speech. However, it is very difficult to collect and store a large speech corpora. Furthermore, the quality of synthesis in these systems depend upon the goodness in joining of the natural speech-units. To overcome these problems, Hidden Markov model (HMM) based speech synthesis system (HTS) was proposed by T. Yoshimura et. al (1999). In HTS, speech is represented by spectral, excitation and durational parameters. These parameters are modeled by contextual-dependent HMMs. Due to this parametric representation, it has following advantages:

1. Smooth and natural sounding speech can be synthesized from small amount of speech corpora.
2. The voice characteristics can be changed.

In this paper, we train a HTS system to synthesis speech in Swedish language.

2. Overview of a basic HTS system

Figure 1 shows an architecture of a basic HMM-based speech synthesis system (H. Zen et. al, 2005). A HTS system mainly consists of two parts: 1) training and 2) synthesis

2.1 Training

Extraction of parameters:
The training part consists of extracting parameters from a given speech database and modeling these parameters by contextual HMMs. To extract features from a given speech signal we used the a high quality analysis tool called STRAIGHT (H. Kawahara et. al 1999). This system extracts three kinds of parameters: 1) spectral, 2) excitation, and 3) aperiodic measures from a speech signal. In this work, we used the Mel-cepstral coefficients (MCEPs) and fundamental frequency (F0) as spectral and excitational parameters respectively. Figure 2 shows the procedure to extract MCEPs from a speech signal.

To extract the fundamental frequency (F0), we performed the voting between the outputs of 1) an instantaneous-frequency-amplitude-spectrum based algorithm, 2) a fixed point analysis called TEMPO from STRAIGHT, and 3) the ESPS get-F0 tool. The aperiodicity measures were calculated as the ratio between the lower and upper smooth spectral envelopes, and averaged across five frequency sub-bands (0-1, 1-2, 2-4, 4-6, and 6-8 kHz).
HMM training:
We followed the same training procedure showed in (H. Zen et al., 2005). In the training, all extracted parameters and durations are modeled in single unified framework. For this, multi-stream model structure is used for simultaneous and synchronous modeling of extracted parameters. We used the Hidden-semi markov model (HSMM), which is same as HMMs except it estimates the state duration probabilities in each iteration of the training. As fundamental frequency (F0) values consist of continuous values in voice regions and discrete symbols (zero) in unvoiced regions, a Multi-space density (MSD) distribution is used to model statistics of F0. The following contextual information is considered in the training to build contextual-HMMs.

- phoneme:
  - {preceding, current, succeeding} phoneme
  - position of current phoneme in current syllable
- syllable:
  - number of phonemes at {preceding, current, succeeding} syllable
  - accent and stress of {preceding, current, succeeding} syllable
  - position of current syllable in current word
  - number of {preceding, succeeding} accented and stressed syllables in current phrase
  - number of syllables {from previous, to next} accented and stressed syllable
  - vowel within current syllable
- word:
  - guess at part of speech of {preceding, current, succeeding} word
  - number of syllables in {preceding, current, succeeding} word
  - position of current word in current phrase
  - number of {preceding, succeeding} content words in current phrase
  - number of words {from previous, to next} content word
- phrase:
  - number of syllables in {preceding, current, succeeding} phrase
  - position in major phrase
  - ToBI endtone of current phrase
- utterance:
  - number of syllables in current utterance

2.2 Synthesis
In the synthesis part, first a given text to be synthesized is converted to a context-dependent label sentence and a sentence MSD-HSMM is constructed by concatenating the parameter-tied context-dependent MSD-HSMMs. Secondly, state durations maximizing their probabilities are determined. Thirdly, speech parameters are generated by using speech parameter generation algorithm. Finally, speech waveform is synthesized directly from the generated MCEPs, F0 and aperiodicity measure sequences using the STRAIGHT vocoder.

Table 1: MOS-based evaluation of the HMM-based speech synthesis system

<table>
<thead>
<tr>
<th>System</th>
<th>Naturalness</th>
<th>Intelligibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>HTS</td>
<td>3.8</td>
<td>4.2</td>
</tr>
</tbody>
</table>

3. Experimental evaluation
We used 1000 Swedish sentences for training with an average duration of 3 seconds. The contextual-labels were generated by using RULSYS, which is developed by KTH and has a long history (Carlson et al., 1982). Five state left-to-right MSD-HSMMs without skip paths were used for training. Each state has a single Gaussian probability distribution function (pdf) with a diagonal covariance matrix as the state output pdf and a single Gaussian pdf with a scalar variance as the state duration pdf.

A perceptual evaluation was conducted to assess the performance of the system. Evaluation is based on the Mean opinion score (MOS) concerning the naturalness and the intelligibility on a scale of one to five where one stands for “bad” and five stands for “excellent”. A total of 5 sentences were synthesized by the system. A group of 10 listeners, comprising both speech and non-speech experts were asked to express their opinion for each sentence on a MOS scale. The results of the test are showed in Table 1. The results indicate that the Swedish speech synthesized by the basic HTS system were perceived natural and intelligible.

These results are encouraging since this is a basic HTS system built for the Swedish language and many further improvements are possible. In our future work, we intend to adapt a more sophisticated model for speech reconstruction, a parametric model of voice source to synthesis different voice styles, and include more prosodic properties in order to increase the naturalness of the produced speech.

4. References
A. Hunt, and A. Black (1996) Unit selection in a concatenative speech synthesis system using a large speech database. in proceedings of ICASSP