Human Language Technologies – The Baltic Perspective
A. Tavast et al. (Eds.)

doi:10.3233/978-1-61499-133-5-162

Creation of HMM-based Speech Model for Estonian Text-to-Speech Synthesis

Tõnis NURK[a,1]
[a] *Institute of the Estonian Language, Estonia*

Abstract. The article describes the creation of Hidden Markov Model based speech models for both male and female voice for Estonian text-to-speech synthesis. A brief overview of text-to-speech synthesis process is given, focusing on statistical parametric synthesis in particular. System HTS is employed to generate voice models. The creation of speech corpus of Institute of the Estonian Language is analyzed. The process of adapting Estonian-related training data and linguistic specification to HTS is described, as well as experiments carried out on data from different speakers, subcorpora and linguistic specifications. The findings from speech model evaluation are given and possible courses of action to improve the quality of HMM-based speech models trained are proposed.

Keywords. speech synthesis, hidden Markov models, speech corpus, speech model

Introduction

The automatic conversion of written to spoken language is commonly called 'text-to-speech' or simply 'TTS'. This paper focuses on statistical parametric approach which is based on hidden Markov model-like models. The method has become competitive with established concatenative techniques over the last few years.

The paper describes the creation of HMM-based speech models that can be used in TTS applications for Estonian. The training of these models are carried out using the system HTS (HMM-based Speech Synthesis System) [1]. The models are adapted to modules for Estonian TTS developed in the environment Festival (The Festival Speech Synthesis System) [2].

Text-to-speech synthesis in Estonia is researched and developed under state-funded programmes at Institute of the Estonian Language [3] and Institute of Cybernetics at Tallinn University of Technology [4].

[1] Corresponding Author:: Institute of the Estonian Language, Roosikrantsi 6 Tallinn 10119 Estonia; E-mail: tonis@eki.ee.

1. Text-to-Speech Synthesis

Text-to-speech synthesis is analogue for human reading. The input to synthesis system is text and output is a speech waveform. A TTS system is almost always divided into two main parts [5]. The first of these, speech analysis, converts text into what we call a 'linguistic specification' and the second part uses that specification to generate a waveform.

Table 1. An example list of context factors which could comprise the linguistic specification.

Preceding and following phonemes
Position of segment in syllable
Position of syllable in word & phrase
Position of word in phrase
Stress/accent/length features of current/preceding/following syllables
Distance from stressed/accented syllable
POS of current/preceding/following word
Length of current/preceding/following phrase
End tone of phrase
Length of utterance measured in syllables/words/phrases

The front end is typically language specific, containing several processes to extract variety of linguistic information from input text. The waveform generation component is largely independent of the language (apart from the data it contains or is trained on).

1.1. Linguistic specification

In the synthesis phase, the input is a linguistic specification. This could be as simple as a phoneme sequence, but for better results we need to include more contextual information. In other words, the linguistic specification comprises whatever factors might affect the acoustic realization of the speech sounds making up the utterance. An example list of context factors is given in Table 1 [6].

2. Statistical Parametric Speech Synthesis

We talk about a statistical parametric approach to speech synthesis particularly when we wish to learn speech models from data. The model is parametric because it describes the speech using parameters, rather than stored exemplars. It is statistical because it describes those parameters using statistics (e.g., means and variances of probability density functions) which capture the distribution of parameter values found in the training data [6].

The employment of hidden Markov models in speech synthesis began after the success of the HMM for automatic speech recognition. HMM-based model is not a true representation of real speech but the availability of effective learning algorithms, automatic methods for model complexity control and computationally efficient search algorithms make the HMM a powerful model.

2.1. Architecture of System HTS

The system used for training speech models and generating speech waveforms is HTS (HMM-based Speech Synthesis System). It consists of two main parts – training and synthesis. During the training spectral coefficients (mel-cepstral coefficients [7] and their dynamic features) and excitation (logarithmic fundamental frequency (log F_0) and its dynamic features) parameters are extracted from speech database and modeled by context-dependent HMM-s (phonetic, phonological and prosodic contexts are taken into account) [8]. Each HMM has state duration probability density functions (PDFs) to capture temporal structure of speech [9]. As a result, the system models spectrum, excitation and durations in a unified framework [10].

During the synthesis part, text to be synthesized is converted to a context-dependent label sequence and then an utterance HMM is constructed by concatenating the context-dependent HMMs according to the label sequence. State durations of the utterance HMM are determined based on the state duration PDFs. Speech parameter generation algorithm generates the sequence of spectral and excitation parameters that maximize their output probabilities. Finally, a speech waveform is synthesized directly from the generated spectral and excitation parameters using the speech synthesis filter [11].

2.2. Advantages of Statistical Parametric Speech Synthesis

Most of the advantages of statistical parametric speech synthesis (against unit-selection synthesis) are related to its flexibility due to the statistical modeling process. Although the training process takes long to complete (up to tens of hours regarding the quantity of training data), it is of little relevancy because it happens only once. The footprint of speech model and synthesis engine is small and is suitable to use in devices with low computational performance.

Intelligible speech can be synthesized with models trained on small amount of data (as little as 100 sentences [12]) because HMM-based speech model is stable and can cover acoustic space despite sparseness of training data.

The main advantage of this approach is its flexibility in changing voice characteristics, speaking styles and emotions.

Since the system requires only language specific database and contextual factors to work, it is rather easily adapted to a language.

2.3. Drawbacks of Statistical Parametric Speech Synthesis

Compared to unit selection speech synthesis the biggest drawback of HMM-based speech synthesis is the lower quality of speech. There seem to be three factors that degrade quality, i.e., vocoder (analysis-synthesis system that reproduces speech [13]), acoustic modeling accuracy and over-smoothing [14].

The speech synthesized by the basic HMM-based speech synthesis system sounds *buzzy* since it uses a mel-cepstral vocoder with simple periodic pulse-train or white-noise excitation. To alleviate the problem, high quality vocoders have been integrated that, for example, take into account the aperiodicity of fundamental frequency. Speaking of Estonian, the letters ‘b’, ‘d’ and ‘g’ (short stops) that appear between voiced letters tend to sound voiced rather than unvoiced so they cannot be synthesized correctly by a simple vocoder.

HMMs are useful for describing transitions between states but regarding one specific state the parameters are static. State-output probability depends on the current state and the probability factor for duration decreases exponentially. This does not hold for real speech. Extra models for describing duration have been employed to overcome this problem.

The statistical averaging in the modeling process improves robustness against data sparseness and the use of dynamic-feature constraints in the synthesis process enables the system to generate smooth trajectories. Compared with natural speech the synthesized speech sounds muffled because the generated speech-parameter trajectories are over-smoothed, thus discarding the natural variability of speech.

3. Speech Corpus

Speech corpus is needed only to train the speech model on. Although rather small amount of training data is needed for synthetic speech to sound intelligible, large corpus of high quality improves the quality of synthetic speech noticeably.

In current work the speech corpus of Institute of the Estonian Language is used. The corpus consists of total 17 hours of high quality speech from five speakers. It was created to be used as a database for unit selection speech synthesis [15], therefore it is suitable for training HMM-based speech models on it.

3.1. Linguistic processing

If we want the speech model trained to be usable in TTS applications, the linguistic specification of the speech model must correspond to the capabilities of text analysis modules. Although the transcription is not always 100% correct it is necessary that labels of training data and speech model correspond to linguistic specification of TTS system for the model to perform well.

In current work, text analysis modules for Estonian speech synthesis developed under Festival have been used. Speech corpus has been transcribed according to the output of linguistic processing unit.

4. Creation of speech model

The goal of current work is to create HMM-based speech model for Estonian TTS using HTS. HTS must be adapted to Estonian and the linguistic specification of speech model is to be compatible with text analysis unit. Adapting HTS to Estonian includes defining phonetic and phonological contexts and preparing training data. Speech models are evaluated by utterance waveforms synthesized with the model.

The training demo on the HTS web site [16] is for English. It consists of data files and scripts. Data files include sound files in RAW format and corresponding utterance files with contextual information, file with description of linguistic specification and utterance files for automatic waveform generation.

4.1. Linguistic Specification

Linguistic specification includes phonetic and phonological context factors. In Estonian, there are 26 qualitative segmental phonemes traditionally [17]: 9 vocals *a*, *e*, *i*, *o*, *u*, *õ* (ɤ according to IPA definition), *ä* (æ), *ö* (ø), *ü* (y) and 17 consonants *f*, *h*, *j*, *k*, *l'*, *l*, *m*, *n*, *n'*, *p*, *r*, *s*, *s'*, *š* (ʃ), *t*, *t'*, *v* (*l'*, *n'*, *s'* and *t'* are palatalized). In speech synthesis nasal *ŋ* and semi-vowel *w* are used additionally, as well as different symbols representing silence. Phonological context factors used are the ones proposed in the training demo for English and described in Table 1.

In current work, experiments have been carried out on different linguistic specifications with the number of phonemes ranging from 30 to 70 (mostly by adding or discarding length markers to phonemes). By increasing the number of phonemes, the number of HMMs increases exponentially. Therefore it is needed to find the balance for the model not to be too simplified nor too complex. Different systems use 50 phonemes on average – English 51 [16], Swedish 53 [18] and Catalan 38 phonemes [19]. Better results for Estonian have been achieved with smaller number of phonemes (up to 50) than with larger number (70).

Phonemes are classified to enable HTS to cluster similar HMMs during the model training process [6]. For example, *a*, *e*, *i*, *o*, *u*, *q* (*õ*), *x* (*ä*), *c* (*ö*), *y* (*ü*) form short vowels and *a:*, *e:*, *i:*, *o:*, *u:*, *q:*, *x:*, *c:*, *y:* long vowels respectively.

4.2. Training corpus

For training corpus, a simple rule holds – the more training data, the better results can be achieved. To train an effective model, the training data should be phonetically rich and balanced. Intelligible speech has been produced using models trained on optimized 100-sentence corpus but several hours of high quality training data provides a good quality speech model.

Experiments with Estonian have been carried out on different subcorpora from speakers Tõnu (6 hours of speech in speech corpus), Liisi (6.5 h) and Riina (1 h). Initial experiments were performed on Riina's subcorpus and since it consists of sentences full of rich linguistic information, 150-utterance test corpus for testing models built for Liisi's and Tõnu's voice was constructed from it. The utterances of the test corpus are not included in training data.

4.3. Evaluation of Speech Models

Trained speech models have been evaluated based on synthesized test sentences. As mentioned before, different linguistic specifications and sizes of training data has been used to create comparable models. For example, training corpora of 100, 250, 500, 1000 and 2000 sentences for Liisi's voice were formed.

As expected, the best results were achieved with largest amount of training data. Surprisingly, there was no perceivable quality difference between speech models trained on large corpus (2000 sentences) for using a) qualitative phonemes only (i.e., only 'short' phonemes) or b) short and long phonemes. Since the text analysis module is designed to use long phonemes as well, it is reasonable to use corresponding speech model in a TTS application. On a small corpus (up to 500 sentences) linguistic specification of qualitative phonemes only performed better than long phonemes included.

Quality of the synthesized speech is good when it is intelligible and there are no abnormalities in phoneme realization. If this condition is fulfilled, the next criterion is mistakes made by text analysis unit. Regarding Estonian TTS, mistake made at determining third quantity degree makes the synthesized speech sound unnatural.

By comparing synthesized test sentences it has been estimated that the most important factor on speech model quality is high-quality phonetically rich training corpus. The next decisive factor is the text analysis unit. The linguistic specification, phonetic and phonological context factors, has less (but still noteworthy) effect on speech model quality.

5. Conclusion

Statistical parametric speech synthesis has been demonstrated to be effective in synthesizing acceptable speech and the approach enables models to be combined and adapted, thus not requiring large speech databases for model manipulation. Many techniques have been employed to improve the quality of speech and regarding HMM-based speech synthesis for Estonian, the next step would be implementing a high-quality vocoder to reduce *buzzyness* of synthesized speech. Further development of text analysis modules and optimizing linguistic specification would improve quality as well.

References

[1] K. Tokuda, K. Oura, K. Hashimoto, S. Shiota, H. Zen, J. Yamagishi, T. Toda, T. Nose, S. Sako, A. W. Black, *HMM-based Speech Synthesis System (HTS)*, http://hts.sp.nitech.ac.jp/ [Last visited on June 19th 2012]

[2] A. W. Black, R. Clark, K. Richmond, J. Yamagishi, V. Strom, S. King, *The Festival Speech Synthesis System*, http://www.cstr.ed.ac.uk/projects/festival/ [Last visited on June 19th 2012]

[3] M. Mihkla, I. Hein, I. Kiissel, T. Nurk, L. Piits, *Eesti keele tekst-kõne süntees*, http://www.eki.ee/keeletehnoloogia/projektid/syntees/tks.html [Last visited on June 19th 2012] (in Estonian)

[4] E. Meister, L. Meister, R. Metsvahi, *Audiovisuaalse kõnesünteesi prototüüp*, http://www.phon.ioc.ee/dokuwiki/doku.php?id=projektid:avsyntees:avsyntees.et [Last visited on June 19th 2012] (in Estonian)

[5] P. Taylor, *Text-to-Speech Synthesis,* Cambridge University Press, Cambridge 2009.

[6] S. King, An introduction to statistical parametric speech synthesis, *Sādhanā* **36(5)** (2011), 837-852.

[7] T. Fukada, K. Tokuda, T. Kobayashi, S. Imai, An adaptive algorithm for mel-cepstral analysis of speech, *Proceedings ICASSP-92* **1** (1992), 137-140.

[8] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, K. Tokuda, The HMM-based Speech Synthesis System (HTS) Version 2.0, *Proceedings of SSW6-2007* (2007), 294-299.

[9] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, Duration Modeling in HMM-based Speech Synthesis System, *Proceedings of ICSLP* **2** (1998), 29-32.

[10] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, Simultaneous Modeling of Spectrum, Pitch and Duration in HMM-Based Speech Synthesis, *Proceedings of EUROSPEECH* **5** (1999), 2347-2350.

[11] S. Imai, Cepstral analysis synthesis on the mel frequency scale, *Proceedings of ICASSP-83* (1983), 93-96.

[12] Y. Takamido, K. Tokuda, T. Kitamura, T. Masuko, T. Kobayashi, *A study of relation between speech quality and amount of training data in HMM-based TTS system*, ASJ Spring meeting 2002, 291-292 (in Japanese).

[13] *Vocoder*, http://en.wikipedia.org/wiki/Vocoder [Last visited on June 19th 2012]

[14] H. Zen, K. Tokuda, A. W. Black, Statistical parametric speech synthesis, *Speech Commun* **51(11)** (2009), 1039-1064.

[15] M. Mihkla, L. Piits, T. Nurk, I. Kiissel, Development of a Unit Selection TTS System for Estonian, *Proceedings of the Third Baltic Conference on Human Language Technologies* (2008), 181-187.
[16] *HMM-based Speech Synthesis System (HTS): Speaker dependent training demo for English*, http://hts.sp.nitech.ac.jp/archives/2.2/HTS-demo_CMU-ARCTIC-SLT.tar.bz2 [Last visited on June 19th 2012]
[17] T. Erelt, M. Erelt, K. Ross, *Eesti keele käsiraamat*, Eesti Keele Sihtasutus, Tallinn, 2007 (in Estonian).
[18] A. Lundgren, *HMM-baserad talsyntes. An HMM-based Text-To-Speech System applied to Swedish*, Master thesis, Royal Institute of Technology, Stockholm, 2005 (in Swedish).
[19] *Descàrrega de les veus Festcat*, http://www.talp.cat/festcat/download.php [Last visited on June 19th 2012] (in Catalan)