Evaluation of Finnish Unit Selection and HMM-based Speech Synthesis

Hanna Silen\textsuperscript{1}, Elina Helander\textsuperscript{1}, Jani Nurminen\textsuperscript{2}, Moncef Gabbouj\textsuperscript{3}

\textsuperscript{1}Department of Signal Processing, Tampere University of Technology, Finland
\textsuperscript{2}Nokia Devices R&D, Finland
\textsuperscript{3}Department of Signal Processing, Tampere University of Technology, Finland
hanna.silen@tut.fi, elina.helander@tut.fi, Jani.K.Nurminen@nokia.com, moncef.gabbouj@tut.fi

\textbf{Abstract}

Unit selection and hidden Markov model (HMM) based synthesis have become the dominant techniques in text-to-speech (TTS) research. In this work, we combine HMM-based signal generation with the front end originally designed for unit selection based Finnish TTS and we evaluate the prosody of the output generated by the two synthesis techniques using the same speech database. Furthermore, we study the effect that the training set size has for the prosody and intelligibility in HMM-based synthesis. The results indicate that the HMM-based approach is capable of providing better prosody than unit selection even if the training set size is severely limited. The size of the training set, however, affects the prosodic quality and intelligibility of the HMM-based synthesizer.

\textbf{Index Terms}: speech synthesis, unit selection, hidden Markov models

\section{1. Introduction}

In text-to-speech (TTS) synthesis, the ultimate goal is to be able to intelligently analyze any input text and synthesize it into speech that sounds completely natural. While even the best systems of today are not at the same level as human beings in reading arbitrary texts, the progress achieved during the past decades has been remarkable. Excellent TTS systems have been developed for many languages, and the best systems can already get very close to the quality of natural speech, especially if the domain of input texts is limited. In addition to the improvements in the quality of TTS systems, a major development in the area of speech synthesis has been the paradigm shift from algorithms requiring expert tuning to data driven techniques. The gradual increase in memory sizes and computational power has also enabled the use of even very large data sets. As a consequence of these changes, it has been possible to move from limited-quality formant synthesis to the currently dominant synthesis techniques of unit selection and hidden Markov model (HMM) based speech synthesis.

In unit selection synthesis \cite{1}, the output speech is generated by selecting and concatenating proper clips of recorded speech stored in a database. The unit size can be chosen in many ways. For example, diphones, triphones and syllables have been used as synthesis units, and it is also possible to use variable-sized units (even a full sentence can be chosen as such if found from the database). Provided that the database is large enough and that the selection works perfectly by always selecting proper units having the correct context, the quality of the synthesized speech can reach the quality of recorded speech. Despite this possibility to achieve the best possible quality, the unit selection approach does not offer a silver bullet for speech synthesis. In reality, it is impossible to design and record speech databases that would contain optimal units for the synthesis of any arbitrary input text. Consequently, the quality in unit selection synthesis varies a lot in practice, unless the input texts are restricted e.g. to contain specific sentence structures in a specific domain.

In contrast to unit selection synthesis, a speech database is not available as such during HMM-based speech synthesis. Instead, the synthesis utilizes HMM models that are trained offline using the speech database as the training data. The HMMs model the behavior of speech parameters related to speech spectrum, fundamental frequency (F0), and durations. During synthesis, speech parameter trajectories are generated from HMMs based on maximum likelihood criteria and converted to speech signal using parametric synthesis techniques. The main strength of the HMM-based approach comes from the fact that the statistical modeling is capable of generating reasonable speech parameter tracks even for cases not included as such in the training data (e.g. for a syllable not available in a similar context in the database). On the other hand, HMM synthesizers suffer from quality degradations caused by the parametric signal generation process. However, in recent years, major improvements in the quality of HMM-based synthesis have been achieved \cite{2}.

In \cite{3}, we described the development of a unit selection speech synthesizer for Finnish. Due to the small database (~80 minutes) and its expressive style, there was a compromise between concatenation smoothness and naturalness. In this paper, the same database is used to train an HMM-based speech synthesizer based on the work presented in \cite{4}, while taking advantage of the text processing module used in the unit selection TTS. We compare the prosody of the unit selection synthesizer with the prosody generated by the HMM-based synthesizer using diphone transplantation. In addition, we study how the size of the training database affects the prosody and intelligibility of the HMM-based speech synthesizer.

This paper is organized as follows. Section 2 discusses the development of the Finnish TTS system, covering both unit selection synthesis and HMM-based synthesis. An evaluation of prosody generated using unit selection and HMM-based synthesis is described in Section 3. The effect of the training set size on HMM-based synthesis prosody and intelligibility is studied in Section 4. Section 5 concludes the paper.

\section{2. Development of Finnish TTS systems}

\subsection{2.1. Characteristics of Finnish}

In Finnish orthography, each phoneme corresponds to a certain grapheme with one exception. The phoneme quantity plays a distinctive role which means that replacing a short phoneme...
with the corresponding long one or vice versa can change the meaning of the word. The word stress is fixed and the primary stress is always on the first syllable. Intonation of a sentence is typically falling and even in questions pitch gets lower towards the end of a sentence. Use of a creaky voice at the end of an utterance is a typical phenomenon in Finnish [5]. Due to the creakiness, pitch may be very low at the end of utterances which can cause problems in synthesis if no attention is paid to it. A more detailed description of Finnish phonetics relevant in TTS is given in [3].

2.2. Finnish unit selection TTS system

Previously, a Finnish unit selection TTS system for academic use has been developed [3]. The system contains modules for text processing as well as for unit selection and waveform concatenation. The selection of the diphone-sized units is based on minimizing the overall cost consisting of the target and join costs. Target subcosts employed in the system are unit’s position in a syllable, word, and sentence, stress, and left and right context. Used join subcosts are linear spectral frequencies, F0, and power. The subcost weighting is manually tuned.

2.3. Adapting HMM-based speech synthesis for Finnish

In this paper, the HMM-based speech synthesis system HTS [6, 7] is applied to Finnish synthesis. Since HTS comes without a text processing part, the text processing module of the existing Finnish unit selection system has been modified to parse the text and generate the context dependent labels required in training and synthesis. Employed contextual features are adapted from the features described in [6]. Part-of-speech, accent, and ToBI endtone are considered irrelevant in Finnish and are therefore excluded. Training of the models is done with the HTS version 2.0.1 and synthesis with an HTK-independent tool HTS engine [7].

The employed speech model in HTS consists of the static and dynamic Mel-cepstral coefficients as well as the static and dynamic log F0 values with binary voiced/unvoiced decisions [7]. However, speech typically contains parts that are actually mixtures of both voiced and unvoiced speech. For example creaky voice common in Finnish cannot be properly modeled with purely voiced or unvoiced signal. Mixed excitation models for HTS have been employed e.g. in [8] but this issue is not studied further in this paper.

3. Prosody: unit selection vs. HTS

3.1. Prosody in unit selection and HMM-based TTS

Unit selection synthesis is based on concatenation of the speech segments selected from a relatively large set of recorded database utterances that contain multiple units sharing the same phonetic identity. In each case, the selection procedure aims at finding the best unit combination, with some limitations, among phonetic identity. In each case, the selection procedure aims at database utterances that contain multiple units sharing the same

3.2. Prosody evaluation techniques

Evaluation of prosody and its acceptability in TTS systems is not trivial. Acoustic measures are not reliable, since for example there can be various acceptable F0 contours for a sentence. Since objective criteria can be misleading, we used listening tests to evaluate prosody.

Comparing the prosody generated using HTS and unit selection directly from their output can lead to several problems. For example, can the listeners focus only on relevant issues when the voice quality is very different due to the parametric speech model used in HTS? In addition, can small labeling errors in the unit selection database affect the result? Due to these reasons, we considered two alternative evaluation techniques: pure prosodic (delexicalized) signals and diphone transplantation.

3.2.1. Pure prosodic signals

The idea of using pure prosodic signals in prosody evaluation was introduced in [9]. The pure prosodic signals are obtained by estimating F0, duration and intensity from the original signals. Sinusoids corresponding to first (F0) and second harmonic frequency are used to generate voiced sounds and unvoiced sounds are represented as silence. The first sinusoid is multiplied with the intensity contour and the second harmonic is modified to have an amplitude of e.g. one fourth of the intensity contour.

The advantage of the method is that prosody can be evaluated without the effect of segmental speech quality. However, according to our experience, these signals are very difficult to evaluate for inexperienced listeners. The durations of individual phonemes can not be evaluated with this technique either and in Finnish they play a very meaningful role, for example the relationship between short and long vowels.

3.2.2. Prosody transplantation on diphones

In diphone synthesis, there is usually only one instance of each diphone. According to the predicted prosody (durations and F0), the diphones are modified for example with PSOLA (pitch synchronous overlap and add) technique. This kind of processing introduces some distortion. In addition, prosody prediction is somewhat difficult and unnatural prosody is one reason that makes diphone synthesis sound unnatural.

Prosody transplantation [10] is a technique to generate more natural prosody for diphone voices. Intonation and phone durations are copied e.g. from real speech or unit selection synthesis database to determine the prosody of diphone voice and modifications are carried out. Nevertheless, voice quality that is also a prosodic cue can not be transplanted.
3.3. Evaluation and comparison of unit selection prosody and HTS prosody

In our experiments, we decided to evaluate prosody using diphone transplantation. The prosody of the unit selection voice hanna_us was compared against the prosody given by HTS voice (hanna_hts) built on the same database. In addition, a smaller HTS voice (hanna_hts_100) was also built using 100 utterances of the same database. The triphone coverage was maximized among the database sentences when building hanna_hts_100.

There is no diphone voice available for our database speaker hanna so for prosody transplantation we used a Finnish female diphone voice suo_fb_jj built on Festival [11]. Although diphone voices can be modified to have an arbitrary F0, major modifications usually lead to distortion. The average F0 value of suo_fb_jj was close to hanna so no scaling was necessary. No scaling was done on duration parameters either, although the original speaking rate of suo_fb_jj was somewhat slower than the speaking rate of hanna. The diphone voice had some strange pronunciation effects, for example problems with boundary gemination, but it was verified that it treated the compared approaches in a similar way.

The speech parameters given by hanna_hts and hanna_hts_100 were transplanted to suo_fb_jj. Further, phone durations from the selected units from unit selection voice hanna_us as well as F0 contours from the resulting speech waveform were transplanted to the diphone voice.

An example of prosody (F0 and phone durations) generated by hanna_hts and hanna_hts_100 as well as prosody provided by the unit selection synthesizer is shown in Figure 1. As can be seen, both HMM synthesizers produced a creaky ending while the unit selection voice did not. This is mainly because creaky parts existing in a database are not easy to concatenate and in addition, there is a limited number of them. The prosody provided by hanna_us has a wider F0 range but also some concatenation problems. Smaller HTS has more monotonic F0 than full HTS, but the trend is very similar. There are some differences with phone durations in the favor of full HTS. The effect of slight faltering due to strange durations was sometimes perceived with hanna_hts_100.

The test set consisted of 20 different sentences that were mainly short and long news texts. 12 listeners were asked to compare the prosody produced by hanna_us against the prosody produced by hanna_hts or hanna_hts_100. Each listener evaluated 20 sentence pairs extracted from the test sentence set in a way that there was an equal number (10) of hanna_us–hanna_hts and hanna_us–hanna_hts_100 evaluations in random order. Two sets of sentence pairs were used in evaluation. Sentences used in hanna_us–hanna_hts evaluations in the first set were used in hanna_us–hanna_hts_100 evaluations in the second set and vice versa. Each listener evaluated one set.

Mean preference and standard deviation values of the listening test are shown in Table 1. Voice hanna_hts was preferred over hanna_us in 92.5% of the sentences. In evaluation hanna_us–hanna_hts_100, HTS was preferred in 82.5% sentences. There was no significant difference between short and long sentences in either case. Based on the results, full HTS voice was rated higher than the smaller voice with a p value of 0.027. With long sentences hanna_hts outperformed hanna_hts_100 (p = 0.017) but for short sentences the difference was not significant.

4. Study on the effects of training set size

4.1. Prosody

We examined prosodic differences between samples generated using HTS trained with full database and HTS trained on 100 sentences. The models were the same as trained in Section 3.3. Because now there was only one synthesis technique included in the test, the prosody comparison was done directly with output speech waveforms produced by hts_engine. Each listener evaluated 10 out of the 20 sentence pairs of the evaluation. For 81.7% of the sentences, the full HTS voice was preferred. Some examples are provided in [12].

4.2. Intelligibility

In speech synthesis, it is highly important to ensure that the output is intelligible. There are three main types of intelligibility tests (e.g. [13]): articulation tests, rhyme tests and speech inference tests.

In this paper, we chose to carry out an articulation test. In the context of synthesis evaluation, articulation can be defined as the ratio of synthesized vocal sounds to those correctly received by the listener. In the test, isolated nonsense words were inserted into Finnish sentences. Several different nonsense words were used and all of them were designed to sound like Finnish words without actually meaning anything. The place in the sentence was selected in such a way that the nonsense word could be any noun. The listeners participating in the test were asked to write down the nonsense word they hear.

The results were analyzed in terms of sound articulation. In each test word, the number of correctly received phonemes was counted. For example, for a nonsense word kemppo, words tempo and kemppa produce an error of one phone while the word temppa an error of two phonemes. The percentage of correctly received words for each voice is shown in Table 2. Mean and standard deviation of correctly received words of six sentences are presented in parenthesis. The number of correctly received words was clearly dependent on the size of the database. All or nearly all of the sentences of hanna_hts and hanna_hts_100 were received with one phone error at the maximum. However, in hanna_hts_30, more than half of the words were received with more than one phone error. Some examples are provided in [12].
5. Conclusions

In this paper, we have discussed issues that are relevant in unit selection synthesis and HMM-based synthesis in the context of Finnish TTS. We have also evaluated the quality of the prosody generated using these two synthesis approaches. Moreover, we have studied the effects that the size of the training database has on the quality of HMM-based synthesis. The results indicate that the HMM-based approach is capable of providing better prosody for Finnish sentences than the unit selection approach even if the size of the training set is limited. The size of the training database, however, affects the prosodic quality and intelligibility of the HMM-based synthesizer. Finally, it should be noted that the results are database specific. With larger databases, it is likely that the prosody of the unit selection approach would be significantly improved.

6. Acknowledgements

This work was partially supported by the Academy of Finland, project No. 5213462 (Finnish Centre of Excellence program 2006-2011).

7. References


