Synthesis of fast speech with interpolation of adapted HSMMs and its evaluation by blind and sighted listeners

Michael Pucher[1], *Dietmar Schabus*[1], *Junichi Yamagishi*[2]

[1]Telecommunications Research Center Vienna (FTW), Austria
[2]The Centre for Speech Technology Research (CSTR), University of Edinburgh, UK
pucher@ftw.at, schabus@ftw.at, jyamagis@inf.ed.ac.uk

Abstract

In this paper we evaluate a method for generating synthetic speech at high speaking rates based on the interpolation of hidden semi-Markov models (HSMMs) trained on speech data recorded at normal and fast speaking rates. The subjective evaluation was carried out with both blind listeners, who are used to very fast speaking rates, and sighted listeners. We show that we can achieve a better intelligibility rate and higher voice quality with this method compared to standard HSMM-based duration modeling. We also evaluate duration modeling with the interpolation of all the acoustic features including not only duration but also spectral and F0 models. An analysis of the mean squared error (MSE) of standard HSMM-based duration modeling for fast speech identifies problematic linguistic contexts for duration modeling.

Index Terms: speech synthesis, fast speech, hidden semi-Markov model

1. Introduction

It is well known that synthetic speech at very high speaking rates is frequently used by blind users to increase the amount of presented information. In data-driven approaches, however, this may lead to a severe degradation of synthetic speech quality, especially at very fast speaking rates. The standard HSMM-based duration modeling [1] is already able to model certain non-linearities between normal and fast speech units since it uses explicit state duration distributions and can thereby take the duration variance of units into account. But for very fast speaking rates this is not sufficient. We therefore propose a duration control method using a model interpolation technique, where we can continuously interpolate HSMMs for normal and fast speaking rate. The HSMMs for fast speaking rate are adapted from HSMMs for normal speaking rate. In addition to interpolation between normal and fast speaking rate, we can also use extrapolation between models to achieve very fast speaking rates that go beyond the recorded original speaking rates. A conventional study [2] already showed that an HMM-based synthesizer with interpolated duration models can outperform a synthesizer with rule based duration model. Their models were, however, based on the so called Hayashi's quantification method I and were theoretically different from our methods that are based on HSMM interpolation and adaptation techniques, which are available from the HTS toolkit today [3].

Some studies have shown that the complex duration changes between normal and fast speech are present at several linguistic levels [4]. Therefore we employ context-dependent linear regression functions for the HSMM duration adaptation to model the duration changes at different linguistic levels. The contexts we used also include high-level linguistic features such as syllable information, phrase information etc. The use of HSMM duration adaptation has another advantage. It makes online processing of the proposed duration control technique possible since normal and fast duration models have the same tying structure and we can straightforwardly perform the interpolation online. This also makes the analysis of the modeling error of standard duration modeling for fast durations easier.

Table 1: Three duration modeling methods used in our evaluation.

Method	Description
SPO–SPO+F	Interpolation between SPO voice and SPO voice with fast duration model.
SPO–SPF	Interpolation between SPO voice and SPO voice with fast duration, spectrum, and F0 model.
SPO	HMM-based duration modeling using acceleration coefficient ρ.

For the evaluation we carried out a comprehension and pair wise comparison test with both blind and sighted listeners. We confirmed that both groups of listeners preferred sentences generated with our method than the conventional method. The proposed method could also achieve lower word error rates (WER) in the comprehension test. The blind listeners were especially good in understanding sentences at fast speaking rates (8-9 syllables per second) compared to non-blind listeners.

2. HSMM Duration Modeling

2.1. Duration modeling methods

All synthetic voices used are built using the framework of a speaker adaptive HMM-based speech synthesis system. Detailed description of the system is given in [5]. Note that our model adaptation is a two-step approach: the first adaptation is for speaker transformation and the second adaptation is for speaking rate adaptation. First we trained an average voice model using several background speakers from speech data at normal speaking rate [6]. We then adapted the average voice model to two Austrian German male speakers (SPO, HPO) using speech data having the normal speaking rates. In the same way we also trained adapted models from speech data with fast speaking rate. We call the adapted models for the fast speaking rates SPF and HPF, respectively. As adaptation data we used a phonetically balanced corpus consisting of approximately 300 sentences for normal and fast speaking rate uttered by each

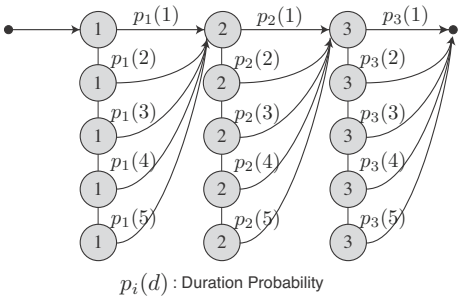

Figure 1: WFST-like illustration of duration models used for TTS systems. Duration probabilities p_i are also transformed to target speakers during speaker adaptation.

speaker. Table 1 shows the three methods that were used in the evaluation. SPO–SPO+F and SPO–SPF are the proposed methods that use interpolations of adapted HSMMs.

To make the differences between the three methods clearer, we explain the temporal structure of the HSMMs [7] and its adaptation. In addition to observations such as the mel-cepstrum and fundamental frequency, each semi-Markov state has a stack of states with associated duration probabilities p_i, illustrated in Figure 1. The duration probabilities p_i are characterized by Gaussian pdfs, and the mean and variance of the pdfs

$$p_i(d) = \mathcal{N}(d; \mu_i, \sigma_i^2). \tag{1}$$

In the HSMM-based parameter generation [1], we use the mean sequence $(\mu_1, \cdots, \mu_N)$ of the Gaussian pdfs corresponding to a given input unit sequence as the most likely sequence. Here N represents the number of states. The easiest and simplest way to control duration is to manipulate the mean of each state using the variance of the state

$$\hat{\mu}_i = \mu_i + \rho\sigma_i^2 \tag{2}$$

and to use a sequence $(\hat{\mu}_1, \cdots, \hat{\mu}_N)$ as a state sequence for the parameter generation. Here ρ is an acceleration coefficient and $\rho > 0$ makes synthetic speech slower and $\rho < 0$ makes synthetic speech faster.

Another way is to transform model parameters for the Gaussian pdfs using a small amount of data for fast speech. There are several possible ways for the transformation and here we employ the CMLLR transform [5], which is given by

$$\mu_i^{\text{fast}} = A_i\mu_i + B_i, \tag{3}$$

$$\sigma_i^{2\ \text{fast}} = A_i^2\sigma_i^2. \tag{4}$$

Here linear regression coefficients A_i and B_i are context-dependent and they are tied through context decision trees having a lot of linguistic questions. To produce speech at various speaking rates, the adapted mean vectors are further interpolated with the original mean vectors

$$\tilde{\mu}_i = (1-w)\mu_i + w\mu_i^{\text{fast}} \tag{5}$$

$$= (1 - w + wA_i)\mu_i + wB_i, \tag{6}$$

where w is the interpolation ratio to control the speaking rate. This interpolation is performed along the state-dependent linear functions obtained from the normal and fast speech. Then a sequence $(\tilde{\mu}_1, \cdots, \tilde{\mu}_N)$ is used as a state sequence for the parameter generation. The same idea may use other acoustic features and system SPO–SPF uses this idea for all the features (spectrum, F0, and duration). System SPO–SPO+F uses this interpolation only for duration pdfs.

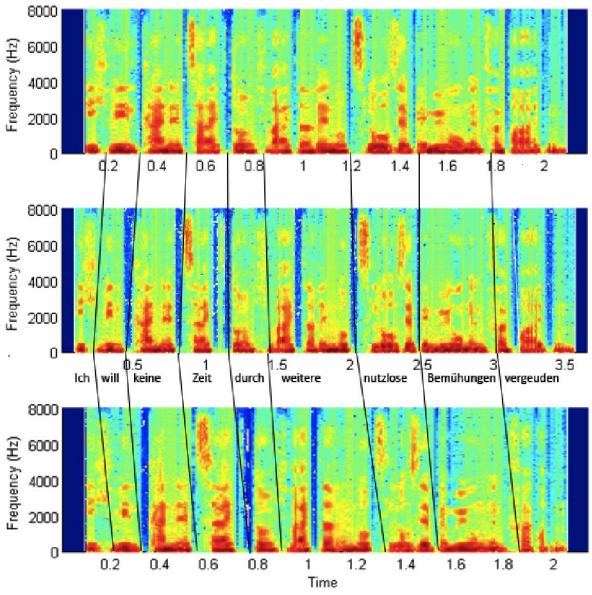

Figure 2: Spectrum of fast sentence with duration interpolation (SPO–SPO+F, top), normal duration (SPO, $\rho = 0$, middle), and fast duration (SPO, $\rho < 0$, bottom).

Figure 2 shows spectra for the utterance "Ich will keine Zeit durch weitere nutzlose Bemühungen vergeuden (I do not want to spend time on additional useless efforts.)". Especially the last word is squeezed with standard duration modeling of fast speech (SPO), which makes it hardly audible. With interpolation this word is much better modeled (top image). Through interpolation we can achieve a better non-linear modeling of duration since we take into account the duration changes from normal to fast speech for contextually modeled units. Speech samples for all three methods can be found on [8].

2.2. Comparison of adapted duration models

It is important for us to analyze the differences between duration values generated by (2) and (6). For such an analysis, the acceleration coefficient ρ_k that is necessary to change from duration μ_k (normal duration) to duration μ_k^{fast} (fast duration) may be calculated by

$$\rho_k = \frac{\mu_k^{\text{fast}} - \mu_k}{\sigma_k^2}. \tag{7}$$

Using (7) , we can define the mean-squared-error of model k that would be produced by using this acceleration coefficient for all models (leaf nodes in the duration clustering tree) by comparing fast duration models and durations produced with ρ_k from normal duration models as follows:

$$e_k = \frac{1}{M}\sum_{i=1}^{M}((\mu_i + \rho_k\sigma_i^2) - \mu_i^{\text{fast}})^2. \tag{8}$$

This value tells us something about the duration errors that a model produces and thereby about the quality that we have achieved in modeling that specific context. Furthermore we define the error for each non-terminal node n as the average error

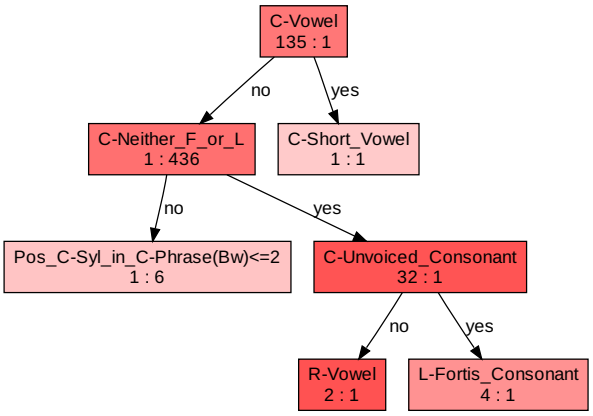

Figure 3: Top of the duration clustering tree. Nodes filled in a darker shade of red have greater average error defined by Equation 9.

of all leaves under n:

$$\bar{e}_n = \frac{\sum_{k \in \text{leaves}(n)} e_k}{|\text{leaves}(n)|}. \quad (9)$$

Figure 3 shows the top of the duration clustering tree with the nodes colored according to their error (Equation 9) for speaker SPO. Looking at the entire tree (which has 1897 leaves and hence 1896 inner nodes) reveals that the subtree rooted at the node labeled "R-Vowel" ("right phoneme is a vowel") has particularly many problematic models. Each node in the figure is labeled with the corresponding question as well as the ratio of the errors of its children, $e_{i \to no} : e_{i \to yes}$. For example, the root node question "C-Vowel" asks whether the central phone is a vowel. We see that the average error for non-vowels is 135 times as big as for vowels. Among the non-vowels, phones which belong to the class "C-Neither_F_or_L" ("central phone is neither fortis nor lenis") are particularly error-prone, and of these, current unvoiced consonants ("C-Unvoiced_Consonant") are not quite as bad.

During the construction of the tree, the class "C-Neither_F_or_L" was defined as containing the phones /l/, /m/, /n/, /ŋ/ and /h/. Of these, only /h/ belongs to the unvoiced consonants, hence the central phone for all models under the node labeled "R-vowel" must be one of /l/, /m/, /n/, /ŋ/. We can see how bad this subtree really is by looking at the cumulative error made by all its leaves (without averaging): The subtree rooted at "R-Vowel" accounts for more than 98% of the total error in the whole tree, but it only accounts for about 13% of the number of leaves.

For speaker HPO, we do not see such clearly distinguished subtrees, however the general trend of consonants having greater average error is confirmed also here. This could be due to more inconsistency of speaker HPO in terms of duration.

3. Evaluation

We evaluated two different male speaker's voices namely SPO and HPO. The duration modeling methods for one voice are described in Table 1. We generated utterances with 7 different durations using the standard HSMM-based synthesis duration method (SPO), interpolation between normal and fast duration model (SPO–SPO+F), and interpolation between normal and fast duration, spectrum, and F0 model (SPO–SPF). In the pair wise comparison we compare the same utterance with the same duration and speaker using different modeling methods.

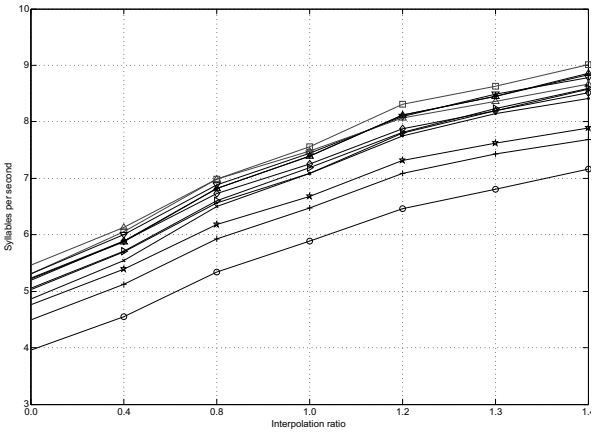

Figure 4: Syllables per second for SPO sentences.

Table 2: Overall word-error rate (WER) and sentence-error-rate (SER) for blind and sighted listeners.

Listeners	WER / SER in %	# sent.
Blind listeners	19.5 / 57.4	108
Sighted listeners	24.4 / 70.4	216

For the evaluation we had 18 sighted listeners (24 to 55 years; 9 female / 9 male) and 9 blind listeners (28 to 56 years; 4 female / 5 male). Within the group of blind listeners we had 2 visually impaired listeners. The users first had to listen once to 12 sentences and write down what they have heard (comprehension test). Afterwards they had to listen to pairs of sentences and decide which sentence they prefer in terms of overall quality. In the pair wise comparison each pair was listened to at least two times by some user.

3.1. Duration of prompts

The duration of prompts is determined by the interpolation ratio and therefore depends on the speaker's duration model. As interpolation ratio we used the following values $[0.0, 0.4, 0.8, 1.0, 1.2, 1.3, 1.4]$. With 0.0 and 1.0 no interpolation is done and only the normal or fast duration model is used. $[1.2, 1.3, 1.4]$ are extrapolation ratios to achieve very fast speaking rates. For the evaluation we had 12 different prompts. Figure 4 shows how many syllables per second are realized for the different sentences by speaker SPO. The fastest sentences contain up to 9 syllables per second.

3.2. Comprehension

Table 2 shows the error rates for blind and sighted listeners. Blind listeners are better in understanding fast speech than sighted listeners. This can also be seen from Figure 5 where we plotted the word-error-rates for the different interpolation ratios. One can see that blind listeners are especially good at recognizing fast speech $[1.3, 1.4]$ where the error rate for sighted listeners is much higher. While the WERs for sighted listeners are almost monotonically increasing, WERs for blind listeners are flat from 1.2 to 1.4. Table 3 shows that in general we can also achieve lower WERs using our interpolation method SPO–SPO+F compared to the standard method.

Table 3: Word-error rate (WER) and sentence-error-rate (SER) per method.

Method	WER / SER in %	# sent.
SPO–SPO+F	16.4 / 61.1	54
SPO–SPF	21.1 / 66.7	54
SPO	23.2 / 66.7	54
HPO–HPO+F	24.3 / 66.7	54
HPO–HPF	25.3 / 63.0	54
HPO	26.3 / 72.2	54

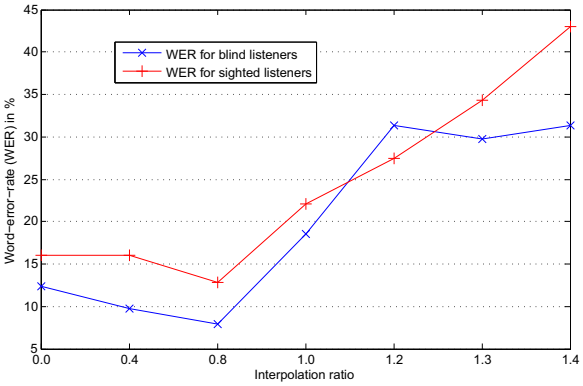

Figure 5: Word-error-rate for blind and sighted listeners per interpolation ratio.

Other work [9] has shown that the error rate for non-blind listeners increases fast from a rate of 10.5 syllables per seconds onwards. In that study, comprehension was subjectively measured by asking listeners how much they could understand of the text. As shown by Figure 4 and 5, the division between blind and sighted listeners concerning understanding can already be seen at around 8 syllables per second when using the objective word error rate measure.

3.3. Pair wise comparison

Figure 6 shows the preference rates for the different methods for all listeners. We see that the adaptive interpolation method where just the duration model is interpolated outperforms the other methods. SPO–SPO+F and HPO–HPO+F are significantly different from the other two methods ($p < 0.05$). The difference is smaller for the HPO voices, since these voices are of a general lower quality than the SPO voices. This lower quality makes the subtle differences of duration modeling more difficult to perceive.

4. Conclusion and future work

We have presented a HSMM-based method for the synthesis of fast speech that outperforms other standard methods on understanding and overall quality. Especially for blind users it is important to have high quality synthesis techniques for fast speech. The adaptive method that we presented can be used with limited amounts of fast speech adaptation data.

In future work we want to investigate the duration rates that are used by blind users of speech synthesis in their everyday use. Furthermore we want to analyze the error of duration modeling on the basis of corpora not only with a comparison of complete models. We also want to investigate the use of fast speech

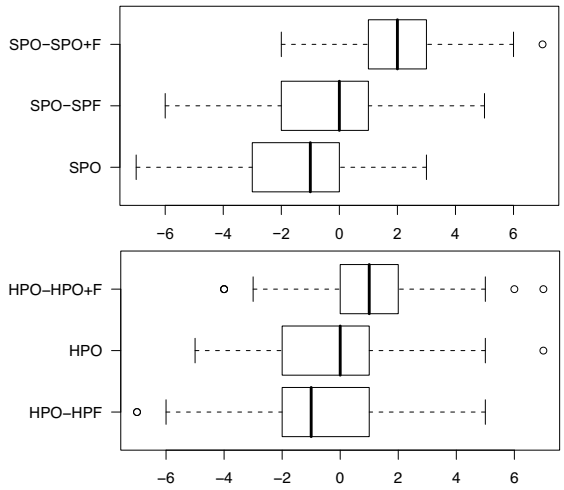

Figure 6: Overall pair wise comparison results per method for speaker SPO (top) and HPO (bottom).

background models and the use of fast duration models from one speaker for another speaker.

5. Acknowledgements

This work was partly funded by the Vienna Science and Technology Fund (WWTF). The Telecommunications Research Center Vienna (FTW) is supported by the Austrian Government and the City of Vienna within the competence center program COMET. Junichi Yamagishi is funded by EPSRC and EC FP7 collaborative projects (EMIME and LISTA).

6. References

[1] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, "A hidden semi-Markov model-based speech synthesis system," IEICE Trans. Inf. & Syst., E90-D, 5, pp.825–834, May 2007

[2] K. Iwano, M. Yamada, T. Togawa, and S. Furui, "Speech-rate-variable HMM-based Japanese TTS system", in Proc. TTS2002, Santa Monica, USA, 2002.

[3] H. Zen, K. Oura, T. Nose, J. Yamagishi, S. Sako, T. Toda, T. Masuko, A.W. Black, and K. Tokuda, "Recent development of the HMM-based speech synthesis system (HTS)," Proc. 2009 Asia-Pacific Signal and Information Processing Association (APSIPA), pp.121–130, Oct. 2009.

[4] E. Janse, "Word perception in fast speech: artificially time-compressed vs. naturally produced fast speech", Speech Communication 42:155–173, 2004.

[5] J. Yamagishi, T. Nose, H. Zen, Z.-H. Ling, T. Toda, K. Tokuda, S. King, S. Renals, "A robust speaker-adaptive HMM-based text-to-speech synthesis", IEEE Trans. Speech Audio Lang. Process. 17 -(6):12081230 Aug 2009.

[6] M. Pucher, D. Schabus, J. Yamagishi, F. Neubarth, V. Strom, "Modeling and interpolation of Austrian German and Viennese dialect in HMM-based speech synthesis" Speech Communication, 52(2):164-179, 2010.

[7] S. Z. Yu, "Hidden semi-Markov models", Artificial Intelligence, 174(2):215-243, 2010.

[8] Samples of fast synthetic speech, `https://portal.ftw.at/projects/vsds/work/publications/fast/fast_synthesis`.

[9] J. Trouvain, "On the comprehension of extremely fast synthetic speech" Saarland working papers in linguistics (SWPL), 1:5-13, 2007.