Large-scale Subjective Evaluations of Speech Rate Control Methods for HMM-based Speech Synthesizers

Tsuneo Kato[1]*, Makoto Yamada*[1]*, Nobuyuki Nishizawa*[1]*, Keiichiro Oura*[2]*, Keiichi Tokuda*[2]

[1]KDDI R&D Laboratories Inc., Japan
[2]Department of Computer Science, Nagoya Institute of Technology, Japan
{tkato, ma-yamada, no-nishizawa}@kddilabs.jp, uratec@sp.nitech.ac.jp, tokuda@nitech.ac.jp

Abstract

Three speech rate control methods for HMM-based speech synthesis were compared by large-scale subjective evaluations. The methods are 1) synthesizing speech sounds based on HMMs trained from corpora at a target speech rate, 2) stretching or shrinking utterance durations proportionally in waveform generation, and 3) determining state durations based on ML criterion under a restriction of utterance duration. The results indicated that the proportional shrinking had significant advantages for fast rate, whereas HMMs trained from slow speech sounds had a slight advantage for slow rate. We also found an advantage of proportionally shrunk speech from a synthesizer trained from slow speech corpora.
Index Terms: HMM-based speech synthesis, speech rate control, subjective evaluation

1. Introduction

Speech rate control is an important issue in speech synthesis, practically. For example, vision-impaired persons tend to choose quite fast speech sounds for the output of TTS (text-to-speech) systems like screen-readers [1] [2]. By contrast, slow sounds are generally more intelligible for elder or hearing-impaired persons [3]. Various methods for speech rate control have been examined separately in these works for both recorded human speech sounds and synthesized sounds. The methods include synthesizing fast speech sounds using acoustic models trained by a corpus of fast speech sound set on the HMM (Hidden Markov Model)-based TTS system [1], controlling the speech rate by stretching or shrinking waveforms in a unit of estimated pitch intervals [3] and time-scale modification by a weighted sum of the adjacent fixed-length segments in consideration of their cross-correlation maximization [4].

Thus, in this paper, we subjectively evaluated three of the speech rate control methods for HMM-based speech synthesis in a large scale experiments with 51 subjects to serve as an example of the quantitative comparison of subjective qualities of speech rate control methods. Our basic TTS system employs HMM-based method [5] which parameterizes acoustic characteristics compactly into HMMs to make it run on cellular phones which have a tight limitation on the use of memory and storage. The evaluated methods were, in short, a corpus-based HMM modeling, proportional modification of the frame-shift interval and ML-based state duration determination under the condition of a total duration. Intelligibility test for a measurement of the clarity and mean opinion scores (MOS) test for preference were conducted.

The remainder is organized as follows. Section 2 describes the specifications of our speech synthesizer and speech rate control methods evaluated in this paper. Section 3 gives the configuration of the evaluations and Section 4 reports the results. Section 5 evaluates speech synthesizers trained only from the slow speech sounds. Our conclusion is presented in Section 6.

2. HMM-based speech synthesizer with speech rate control

In this section, we introduce a duration modeling built in our speech synthesizer first since the formulation of the duration modeling is directly related to one of the speech rate control methods, namely the ML-based method in Section 2.2. Then, we describe each of the three methods in detail.

2.1. Duration modeling in the target speech synthesizer

In our system, the state duration is directly modeled by a Gaussian distribution, not by a state transition probability. Thus, strictly speaking, the speech synthesizer is based on Hidden Semi Markov Model (HSMM) rather than HMM. This state duration modeling is a current standard method which has been introduced to HTS [6]. As the duration control was not enough with only the state duration modeling, phone duration modeling [7] to stabilize phone duration is introduced to the acoustic models in addition to the state duration modeling. Based on the maximizing likelihood (ML) criterion of the model function which is the weighted sum of the likelihoods of the state duration and of the phone duration, the state durations $d^*_{n,s}$ of the s-th state of phone n are derived as

$$\begin{aligned} d^*_{n,s} &= \arg\max_{d_{n,s}} [\log p_{n,s}(d_{n,s}) + w \cdot \log p_n(\sum_{i=1}^{S} d_{n,i})] \\ &= m_{n,s} + \frac{m_n - \sum_{i=1}^{S} m_{n,i}}{\frac{1}{w} \cdot \sigma_n^2 + \sum_{i=1}^{S} \sigma_{n,i}^2} \cdot \sigma_{n,s}^2 \end{aligned} \quad (1)$$

where $m_{n,s}$, $\sigma^2_{n,s}$, m_n and σ^2_n are the mean and variance of the duration of the s-th state of phone n and those of phone n, respectively. w is a likelihood weight for the phone duration models to the state duration models and S is the number of states defined in a phone. It should be noted that the numerator of Eq. (1) is not zero, because the samples for estimating distributions of state duration models of a phone are not identical to those of the correspoinding phone duration model due to independent state tying and tying of phones. It is also noted that the conventional duration modeling without the phone duration model corresponds to the case when $d^*_{n,s}$ equals to $m_{n,s}$. In this study, based on the preliminary investigation, w is set infinite so that the phone durations are determined only by the phone duration models.

2.2. Speech rate control methods

In this study, the following three speech rate control methods were examined:

1. **Corpus-based**: Synthesizing speech sounds based on a set of HMMs trained from the target speech rate corpus, in which every utterance is pronounced at an adequate speech rate by the same narrator.
2. **Proportional Stretch/Shrink**: Changing the frame-shift interval T (default value of T is 5 ms) in waveform generation for the target duration of the utterance proportionally. The HMMs used here are trained from a normal rate speech corpus.
3. **ML-based**: ML-based determination of the state durations under a restriction of the total duration of the utterance. Given the total duration of the utterance, ML criterion derives the state duration of each state by multiplying the second term of the right-hand side of Eq. (1) by a single constant k as

$$d^*_{n,s} = m_{n,s} + k \cdot \frac{m_n - \sum_{i=1}^{S} m_{n,i}}{\frac{1}{w} \cdot \sigma_n^2 + \sum_{i=1}^{S} \sigma_{n,i}^2} \cdot \sigma_{n,s}^2 \quad (2)$$

The greater the variance is, the greater the state duration changes in the speech rate control.

Practically, the values of T and k were adjusted for each stimulus so that the total durations of speech sounds generated by each method would be the same as the synthesized speech by the Corpus-based method.

Of course, the Corpus-based method is usually the most costly since it is not easy to record speech sounds and annotate them for building the corpus.

3. Configurations of Evaluations

3.1. Acoustic modeling

The acoustic models are trained by HTS version 2.1 with speech corpora separated by two narrators and the speech rate. Each corpus consists of utterances of approximately 400 sentences. The sentences consist of the ATR phonetically balanced 503 Japanese sentences (amounting to 35 ˜40 minutes). Speech sounds for the training were down-sampled to 16 kHz. The corpora of each narrator consist of three sets, a set of normal speed utterances used for the speech rate control methods, Proportional Stretch/Shrink and ML-based, and two sets of both fast and slow speed speech sounds for the Corpus-based method. The acoustic features are the same as the preceding work by Zen et al. at Blizzard Challenge 2005 [8], 39-order mel-cepstrum including the 0th-order coefficient. Five-state HMMs were trained by Baum-Welch reestimation method. The average speech rates of the synthesized stimuli by the Corpus-based method are shown in Table 1.

Table 1: *Average speech rates of the speech corpora in morae per second.*

	Fast	Normal	Slow
Female	10.2	7.7	5.3
Male	9.3	7.8	5.8

3.2. Evaluations and subjects

We set up two types of evaluations, the intelligibility test and the mean opinion scores (MOS) test. In the intelligibility test, the subjects listened to synthetic sounds and wrote them down. The synthetic sounds consisted of grammatical but meaningless Japanese sentences [9]. The length of each sentence was approximately 16 morae. In the MOS test, the subjects also listened to speech sounds and scored them on a 5-point category-scale (1: "Very poor", 2: "Poor", 3: "Fair", 4: "Good", 5: "Very good"). The stimuli contained synthesized sounds of the typical contents such as newspaper articles, tourists' dialogues, etc. These two evaluations were conducted over the three speech rate control methods to the fast and the slow rates. The intelligibility test was done first and then the MOS test was. We also conducted MOS test on the synthetic sounds with and without the phone duration restriction introduced in Section 2.1 for each speech rate by the Corpus-based method as a preliminary experiment.

Table 2: *MOS scores and P-values for the basic HMM-based speech synthesizer with and without phone duration restriction.*

Acoustic models	Phone duration restriction		P-value
	Without	With	
Female Fast	3.20	3.24	0.664
Female Normal	3.54	3.53	0.922
Female Slow	2.96	3.20	0.008
Male Fast	3.52	3.62	0.164
Male Normal	3.84	3.87	0.842
Male Slow	3.04	3.12	0.416

The subjects were fifty-one non-professional normal-hearing people collected to keep in even distribution regarding their age, from their twenties to fifties, and their gender (#female = 25, #male = 26). Each stimulus was presented to both ears through a headphone at 16 kHz sampling. The intelligibility test consisted of one session where all of the stimuli were arranged in random order regarding the narrators, three speech rates and three speech rate control methods. In the MOS test, the evaluation was separated into two sessions by the narrators. The speech rates and the speech rate control methods were arranged randomly in each session.

3.3. Statistical analysis for the results

For testing the differences in scores of each method statistically, we conducted the following nonparametric hypothesis testing scheme with the overall significance level $\alpha = 0.05$. First, Kruskal-Wallis multiple comparison (hereafter the H-test) was conducted to examine whether at least one significant difference exists among all possible pairs of the three speech rate control methods. Then if the difference among the scores was significant by the preceding H-test, the pairwise comparisons by Mann-Whitney U-test were applied to identify the differential pair. The significance levels for each pairwise comparison were adjusted depending on the difference of the rank of the scores by the following formula according to Ryan's procedure [10]

$$\alpha'_r = \frac{2\alpha}{m(r-1)} \quad (3)$$

where m represents the number of groups in the comparison and r represents the step count of the pairwise test. This procedure is processed as follows. The scores of m groups are sorted first. Then the most different pair, the lowest and the highest, is tested with α'_m. If the difference is significant, the step count r is decremented and the second most different pairs, (a) the second lowest and the highest and (b) the lowest and the second highest, are tested. The procedure is terminated when the pairs with $r = 2$ (the pairs of scores next to each other, intuitively) have been tested or the difference is not significant at a previous step. For instance, if the difference of the pair (a) as shown above is not significant, the differences of the pairs, (c) the third lowest and the highest and (d) the second lowest and the second highest, are determined as 'not significant' without testing.

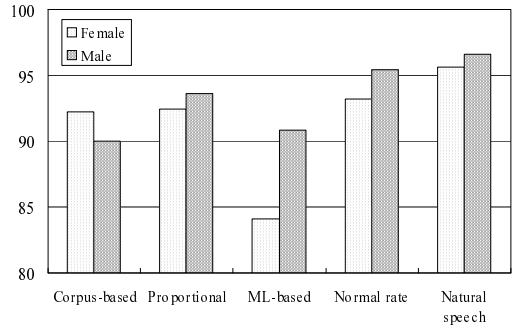

Figure 1: *Syllable hit rates [%] for the fast rate speech sounds.*

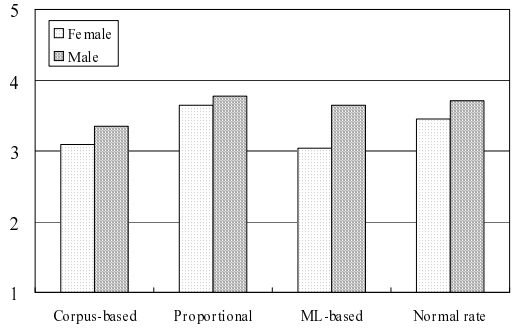

Figure 2: *MOS scores for the fast rate speech sounds.*

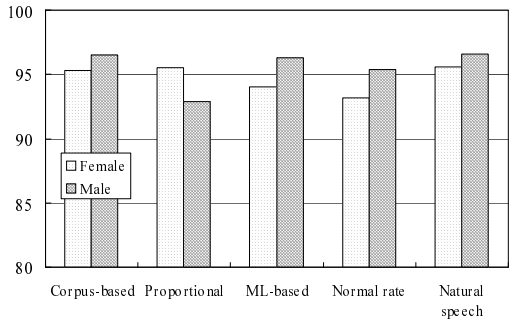

Figure 3: *Syllable hit rates [%] for the slow rate speech sounds.*

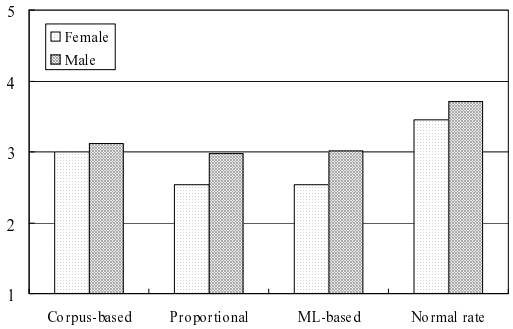

Figure 4: *MOS scores for the slow rate speech sounds.*

4. Experimental Results

4.1. Effect of adoption of phone duration modeling

We first report the results of the preliminary evaluation, which tests the differences of the phone duration restrictions in Section 2.1. The MOS scores and the resulting P-values of the statistical testing of each difference are shown in Table 2 (U-test with $\alpha = 0.05$). P-value is the probability of observing the difference with an assumption that the null-hypothesis is true. In the table, the MOS scores are compared between with and without the restriction on six acoustic models, three speech rates for two narrators. The number of votes for the stimuli is 255 per acoustic model. The MOS of the slow synthetic sounds of the female narrator with the phone duration restrictions was significantly superior to the MOS without the restriction. Meanwhile, the other differences are not significant. Since there was no significant inferiority by the phone duration restriction, we introduced this restriction for all acoustic models for the subsequent evaluations.

4.2. Comparison for fast rate speech

The results of the intelligibility test and the MOS test for the fast rate synthetic speech sounds are illustrated in Figure 1 and Figure 2. The heights of bars show averages of syllable hit rate [%] in Figure 1 and MOS scores in Figure 2, respectively. For reference, the scores of the synthetic sounds and those of the natural (narrators' own) voices both at the normal speech rate are also shown in the figures. The numbers of the evaluated stimuli for each set are 102 sentences for the intelligibility test and 255 for the MOS test.

By the statistical tests described in Section 3.3, in terms of the intelligibilities, the ML-based method is significantly inferior to the other methods for the female narrator and the Proportional Stretch/Shrink method is significantly superior to the Corpus-based method for the male. As for the MOS, the Proportional Stretch/Shrink method is superior to others for the female narrator and the Corpus-based method is inferior to others for the male, both significantly. Generally, the Proportional Stretch/Shrink method seems to realize higher quality for controlling synthetic speech sounds to the fast rate.

We suppose the inferior performance of the Corpus-based method for the fast rate is caused by degradation of articulation by the narrators. We suppose the inferiority of the ML-based method is due to imbalanced state durations caused by the imbalanced variances of the state duration in Eq. (2). Besides, the acoustic models including dynamic features (Δ and $\Delta\Delta$) of only the corpus of normal speech rate are considered harmful.

4.3. Comparison for slow rate speech

The results of the evaluations for the slow synthetic speech sounds are illustrated in Figure 3 and Figure 4. The numbers of the stimuli are the same as in Section 4.2. As a result of the statistical analysis, the intelligibilities are not significantly different among the methods for either narrator. As for the MOS test, the Corpus-based method is significantly superior to the others for the slow synthetic sounds of the female narrator. For the male narrator, no significant difference is observed. In the slow rate speech control, there seems to be a trade-off between the cost of constructing the slow speech corpora for the Corpus-based method and the preference of the synthesized speech sounds.

5. Evaluations for speech synthesizers trained only from slow speech sounds

Summarizing the results in Section 4, the Proportional Stretch/Shrink method is superior for the fast rate speech sounds and the Corpus-based method has a slight advantage for the

Table 3: *Syllable hit rates [%] of the proportionally shrunk speech sounds from the acoustic models of the normal and slow speech rate.*

Narrator of acoustic models	Target speech rate	Speech rate of acoustic models		P-value
		Normal	Slow	
Female	Fast	89.5	91.1	0.27
	Normal	91.8	95.7	1.6e-3
	Slow	94.5	96.3	0.015
Male	Fast	90.5	92.0	0.077
	Normal	93.7	94.5	0.89
	Slow	94.6	93.8	0.87

Table 4: *MOS scores of the proportionally shrunk speech sounds from the acoustic models of the normal and slow speech rate.*

Narrator of acoustic models	Target speech rate	Speech rate of acoustic models		P-value
		Normal	Slow	
Female	Fast	3.25	2.96	6.7e-5
	Normal	3.58	3.41	0.0028
	Slow	2.43	2.82	3.6e-11
Male	Fast	3.67	3.30	2.1e-8
	Normal	3.73	3.55	0.0032
	Slow	2.68	2.92	7.1e-5

slow rate though it needs an additional cost for constructing slow speech corpus. Thus, in this section, we hypothesize that the synthetic sounds at various speech rates are able to be produced by the Proportional Shrink method on the speech sounds synthesized by an acoustic model trained from a slow speech corpora. We evaluate the speech sounds synthesized by the Corpus-based acoustic models of normal and slow rate speech sounds and speech-rate-controled by the Proportional Stretch/Shrink method.

5.1. Design of the evaluations

We prepared twelve sets of stimuli, using two acoustic models of the slow and normal speech rates both trained in Section 2 for two narrators and controlling the speech rate to fast, normal and slow rates by the Proportional Stretch/Shrink method if the target speech rate and the speech rate of the acoustic models are different. For evaluations, another fifty-one subjects were collected containing 59% of the same subjects as the preceding evaluations in Section 4 with an interval of more than three months. The interval is considered long enough to evaluate even for the same subjects. Other configurations such as the types and the procedure of the evaluations are the same as those described in Section 3. The number of votes for each set of stimuli is 204 for the intelligibility test and 510 for the MOS test.

5.2. Results of the evaluations

Table 3 and Table 4 show the results of intelligibility test and of the MOS test respectively. The differences of the pairs arranged next to each other were statistically tested by the U-test with $\alpha = 0.05$.

According to the results, the normal rate speech sounds of the female narrator synthesized by the slow rate acoustic models were superior to those synthesized by the normal rate acoustic models regarding the intelligibility. It might be caused by the emphasized dynamics of articulations of the slow speech sounds. The MOSs of the speech sounds synthesized by the normal rate acoustic models generally seem superior at the fast and the normal speech rates. However, it might be an option to use the acoustic models trained only from slow speech sounds and control the speech rate by the Proportional Stretch/Shrink method for the purpose of realizing synthetic sounds on various speech rates when the circumstances require higher intelligibility. This is because, in addition to the result of the intelligibility evaluation in this section, the Corpus-based method is superior for synthesizing slow speech sounds as described in Section 4.3.

6. Conclusions

In this paper, we reported the results of the subjective evaluations, the intelligibility test and the MOS test, for three speech rate controlling methods on HMM-based speech synthesis. The results indicated that the Proportional Stretch/Shrink method was superior to other methods including a Corpus-based method for controlling to the fast rate. They also indicated that the Corpus-based method had a slight advantage on the preference for controlling to the slow rate. We also evaluated speech sounds at controlled speech rates by the Proportional Stretch/Shrink method from synthesized speech sounds generated by acoustic models trained only from slow speech sounds and the results indicated that the slow speech synthesizer with speech rate control by the Proportional Stretch/Shrink method would be one option if the intelligibilities of synthesized speech sounds were weighed by circumstances.

7. Acknowledgements

We thank Mr. Naoki Ito and Mr. Shinji Takagi of Nagoya Institute of Technology for their valuable cooperation in synthesizing speech sounds for stimuli.

8. References

[1] T. Nishimoto, S. Sako, S. Sagayama, K. Ohshima, K. Oda and T. Watanabe, "Effect on Learning on Listening to Ultra-Fast Synthesized Speech," Proc. of the 28th IEEE EMBS Annual International Conference, pp.5691-5694, 2006.

[2] C. Asakawa, H. Takagi, S. Ino and T. Ifukube, "Maximum Listening Speeds for The Blind," Proc. of the 2003 International Conference on Auditory Display, pp.ICAD03-276-ICAD03-279, 2003.

[3] S. Seiyama, A. Imai, T. Mishima, T. Takagi and E. Miyasaka, "Development of A High-Quality Real-time Speech Rate Conversion System," Trans. on IEICE, D-II. Vol.J84-D-II. No.6, pp.918-926, 2001 (in Japanese).

[4] R. Suzuki and M. Misaki, "Time-scale modification of speech signals using cross-correlation functions," IEEE Trans. on Consumer Electronics, Vol.38, pp.357-363, 1992.

[5] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," Proc. of Eurospeech 1999, pp.2347-2350, 1999.

[6] HMM-based Speech Synthesis System: HTS, http://hts.sp.nitech.ac.jp/.

[7] Z. Ling, Y. Wu, Y. Wang, L. Qin and R. Wang, "USTC System for Blizzard Challenge 2006 an Improved HMM-based Speech Synthesis Method," Proc. of Blizzard Challenge 2006 workshop, 2006.

[8] H. Zen and H. Toda, "An Overview of Nitech HMM-based Speech Synthesis System for Blizzard Challenge 2005," Proc. of Interspeech 2005, pp.93-96, 2005.

[9] Japan Electronics and Information Technology Industries Association (JEITA), "JEITA IT-4001 Speech Synthesis System Performance Evaluation Methods," 2003 (in Japanese).

[10] T. A. Ryan, "Significance Tests for Multiple Comparisons of Proportions, Variances, and Other Statistics," Psychological Bulletin, 57(4), pp.318-328, 1960.