Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Effect of Learning on Listening to Ultra-Fast Synthesized Speech

Takuya Nishimoto, Shinji Sako, Shigeki Sagayama, Kazue Ohshima, Koichi Oda and Takayuki Watanabe

Abstract— **A text-to-speech synthesizer that would produce easily understandable voices at very fast speaking rates is expected to help persons with visual disability to acquire information effectively with screen reading softwares. We investigated the intelligibility of Japanese Text-to-Speech systems at fast speaking rates, using four-digit random numbers as the vocabulary of the recall test.**

We also studied the fast and intelligible text-to-speech engine, using HMM-based synthesizer with the corpus with fast speaking rate. As the results, the statistical models trained with the fast speaking corpus was effective. The learning effect was significant in the early stage of the trials and the effect sustained for several weeks.

I. INTRODUCTION

According to a survey by the Japanese Cabinet Office in 2001, there are around 306,000 persons with visual disability in Japan. Those who can use Braille among them are less than 10%. Braille is not commonly used because it is especially difficult for elderly people with visual disability to learn it. On the other hand, more than 10,000 persons with visual disability use personal computers (PC), according to the same survey.

It is important for persons with visual disability to use PCs and/or Internet with voice, which allows real-time communication and gives the chance of taking part to social activities. Recently, screen-readers for the Japanese version of Microsoft Windows have been developed, which help the persons with visual disability to access the Web and read or write e-mails. They are sold at low prices and do not require any additional hardware.

To make such systems easy to use, we must consider how quickly and accurately the users understand information by listening to the voice of the Text-to-Speech (TTS) systems. Watanabe [1] investigated how the PC users with visual disability in Japan are setting the voice of screen readers. He reported that most of the users are using TTS with the maximum reading rate that the softwares can read with. In many cases, it was double the speed of the normal speaking rate.

The aim of our research is to improve the voice quality of the TTS for the screen readers. We must therefore evaluate objectively the various factors of TTS. For example, when developing better speaker models for TTS, it is important to prove that the technique is effective. In this paper, we focus on the relationship between the reading rate of TTS and the intelligibility (recall rate). Using the method we propose, one can expect to be able to choose the TTS engine or the speaker model which is the most intelligible, in other words, the one through which users with visual disability can understand information via fast-speaking voice in the shortest time.

There are many related works which propose different evaluation methods of speech synthesis systems' performance. For example, the evaluation method standardized by JEITA [2] consists of tree parts: the correctness of pronunciations, the intelligibility and the overall judgment. Our research, however, focuses on the word intelligibility test. The familiarity of the words used may have an effect on the performance of the recall test. We thus chose to use four-digit random numbers as the vocabulary of the recall test, as one can assume that the familiarity to random numbers does not vary much among the users. Very few works investigated the learning effect for synthesized voices, but the users may adapt to the voice and performance then increase during the sessions of the recall tests. We assumed that this was also the case for users of TTS, and that individual variations of this learning effect could not be disregarded. We thus investigated how the users can learn to listen to fast speaking synthesized voice.

II. RELATED WORKS

A. Listening Speed for the Persons with Visual Disability

In a prior work, Asakawa et al. [3], [4] created rapidly-spoken Japanese speech sentences by using the time-stretch/compression function of the CoolEdit audio processing software and shortened recorded human voices linearly on the time axis. Then the stimuli were evaluated by subjects with visual disability who are skilled users of screen reading softwares. According to their work, the suitable speed and the highest speed are defined respectively as the speaking rate for which listeners can understand a sentence sufficiently, and the speaking rate for which listeners can understand approximately 50% of the words in the sentence. They report results which show that skilled listeners of synthesized speech assess the suitable speed (recall rate = 90%) as 1100-1170 morae/minute and the highest speed (recall rate = 50%) as 1400-1500 morae/minute, respectively. Most of the commercial text-to-speech engines could not produce voices at such fast speaking rates, so their work suggested a way to improve non-visual user interfaces for people with visual disability.

Takuya Nishimoto, Shinji Sako, Shigeki Sagayama are with the Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8656, JAPAN nishi@hil.t.u-tokyo.ac.jp

Kazue Ohshima, Koichi Oda and Takayuki Watanabe are with the Department of Communication, Tokyo Woman's Christian University, 2-6-1 Zenpukuzi, Suginami-ku, Tokyo, 167-8585, JAPAN

B. HMM-based Speech Synthesis

In our study, GalateaTalk [5], [6] and HTS (HMM-Based Speech Synthesis System) [7] are used as the Japanese TTS system. GalateaTalk, the first open-source Japanese TTS, consists of a text analyzer and a waveform generation engine. The text analyzer decomposes arbitrary Japanese input texts containing Kanji, Kana, alphabetic and numeric characters, and optionally embedded tags specifying the speaking style, linguistic information including pronunciation, accent type, part of speech, etc., partly utilizing ChaSen [8] and newly developed dictionaries for Japanese morphological analysis. The waveform generation engine, derived from the corresponding part of HTS, is a HMM(Hidden Markov Models)-based speech synthesizer. The speech spectrum is considered as the observation output of the states of an HMM, in the same way as in speech recognition systems. Additionally, two HMMs are respectively used to model F_0 and the duration of each phoneme, i.e. the number of frames in each state.

We created rapidly-speaking synthesized voices by using the HMM-based speech synthesizer and rapidly-speaking statistical models [9], which are trained from a corpus of rapidly-speaking human voices. We investigated more precisely the durations of phonemes of the statistical models in [12]. The system can produce voices with a speed faster than 1500 morae/minute (i.e. 25 morae/second). The duration of each segment of the phoneme can be generated using the means and variances trained by the corpus. When the target speaking rate is higher than the speaking rate of the voices of the corpus, if the variance of the duration is large, the number of frames corresponding to the segment is highly decreased, and if the variance of the duration is small, it is less decreased. One expects that the use of HMM-based TTS and rapidly-speaking statistical model enables to produce intelligible and fast-speaking TTS.

C. Perception of Speech and Speaker

Legge et. al. [13] investigated the learning of unfamiliar voices. In their experiment, subjects listened to a series of recorded voice samples obtained from unfamiliar speakers and were then given a two-alternative forced-choice recognition test. According to their results, voice learning was inferior to face learning.

Palmeri et. al. [14] suggested that detailed information about a speaker's voice is retained in long-term episodic memory representations of spoken words. In their work, recognition memory for spoken words was investigated with a continuous recognition memory task. Independent variables were the number of intervening words between initial and subsequent presentations of a word, the total number of speakers in the stimulus set, and whether words were repeated by the same voice or a different one.

Nygaard et. al. [15] suggested that speech perception may involve speaker-contingent processes. The results of their experiments showed that the ability to identify a speaker's voice improved intelligibility of novel words produced by that speaker.

TABLE I
DURATION OF THE SYNTHESIZED SPEECH

Speaker	Duration of sentence(ms)	Speed(morae/min)
Normal	2090	402
Fast	1325	634

III. FIRST EXPERIMENT

A. Procedure

We performed a preliminary experiment to show that the recall rates decrease at fast speaking rates, and that subjects can learn to listen to fast speaking voices during the tests [9].

We created two corpora from the voice of a male professional narrator. The normal-speed corpus is a set of speech utterances corresponding to 503 phonetically balanced Japanese sentences. The fast-speed corpus is a set of speech utterances corresponding to 100 sentences which are a subset of the 503 previous sentences.

Both corpora were used to create statistical speaker models for GalateaTalk TTS engine. In the model, each phoneme consists of five states. Table I shows the duration and the speed of the speech synthesized using the speaker models on the Japanese test sentence "Bango-wa ichi ni san yon desu" (The number is one two three four). The duration of each phoneme of this synthesized speech is as follows:

- Normal: b[70] a[95] N[100] g[65] o[100] o[80] w[95] a[100] i[95] ch[105] i[90] n[55] i[95] s[100] a[90] N[100] y[85] o[85] N[100] d[45] e[95] s[165] U[80],
- Fast: b[45] a[70] N[60] g[45] o[45] o[70] w[45] a[50] i[60] ch[65] i[40] n[50] i[60] s[75] a[65] N[65] y[70] o[55] N[65] d[30] e[55] s[80] U[60].

It turned out that the TTS with the fast speaker model could speak approximately 1.6 times faster than the TTS with the normal speaker model. We also confirmed that the duration is quantized with the frame shift time (5ms). The frame is the unit of generation and processing of speech signals inside the waveform generation engine.

We used two configurations of TTS for testing. The Fast-1 uses the statistical information of the Fast corpus to produce the duration, F_0 and the spectrum. The Fast-2 uses the statistical information of the Fast corpus to produce the duration, and the statistical information of the Normal corpus to produce F_0 and the spectrum. Both conditions use the same duration model, therefore the speaking rates of their synthesized voices are exactly the same.

The subjects were three women who are university students and are not visually/hearing impaired.

The stimuli are four-digit random numbers embedded in the sentence as described above. The subjects used Windows PC to listen to the stimuli and input the answer. Headphones (Sony MDR-CD480) were used during the tests. The subjects were indicated that if there were some digits which they could not hear, question marks were available for input instead of the numbers, such as '1 ? 3 0.'

The speaking rates we used were 0.3, 0.4, 0.5, 0.6 and 0.7, these values indicating the ratio of the duration of the

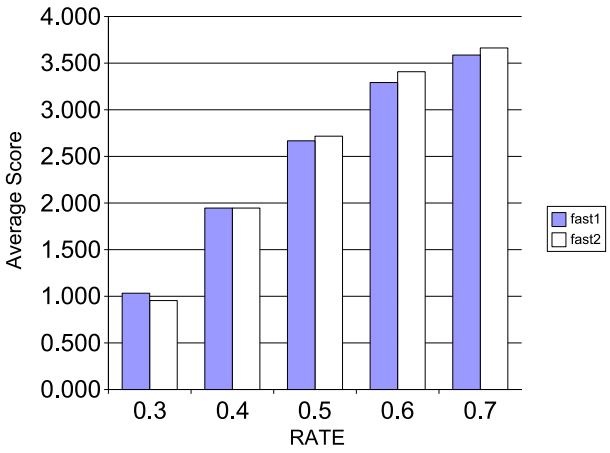

Fig. 1. Average score for each speed in the first experiment.

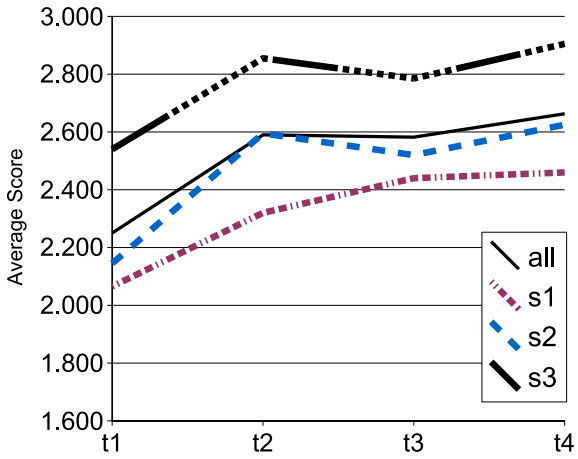

Fig. 2. Learning effect in the first experiment (s1-s3:subjects, all:average).

synthesized speech to the duration of the original speech. The durations of each segment of the phonemes were not generated by proportional allotment, but generated in consideration of the means and the variances of the statistical model.

The trial T1, consisting of 200 recall tests, was performed, followed by the first training session, consisting of 50 training tasks, and the trial T2, which is the same as T1.

During the training task, the subjects listen to a fast-speed voice, then input the answer, and finally listen to the correct answer in the form of the corresponding normal-speed voice.

The trial T3, the second training session and the trial T4 were conducted three days after in the same way.

B. Results

Figure 1 shows the average score of the recall task in the first experiment. The recall rate decreases for fast speaking rates, therefore proving the appropriateness of our experiments. The rates of Fast1 and Fast2 are very similar, showing that the difference of spectrum has little impact on the speech quality if the durations are the same.

Figure 2 shows the learning effect in the first experiment. The results indicate that all the subjects learned to listen to the voice very quickly, and that the effect sustained for at least several days.

IV. Second Experiment

A. Procedure

We performed a second experiment to investigate the individual variations of the learning effect [10], [11].

The experiment was conducted over three weeks. Nineteen female university students participated in the sessions of the first week. Five of them also participated in the sessions of the second and third week succeedingly. The data of sixteen people was used for analysis. These sixteen people were not visually/hearing impaired, and four persons among them participated in all sessions.

The stimuli are four-digit random numbers embedded in the same sentence as in the first experiment. The subjects used Windows PC to listen to the stimuli and input the answer. Headphones (Panasonic RP-H750-S) were used during the tests. Question marks were available for input.

In this experiment, the trial T1, consisting of 200 recall tests, was performed, followed by the first training session, consisting of 25 training tasks, and the trial T2. The trial T1 consisted of the iterations of a set of 50 tasks, and the trial T2 consisted of the iterations of the set of the same tasks in the reverse order.

The waveform generation engine was changed from GalateaTalk to the original HTS, and the frame shift time was changed from 5ms to 2ms. To produce the duration, F_0 and the spectrum, the Fast and the Normal speaker models used the statistical information of the Fast and Normal corpora, respectively.

In order to improve speech quality, we introduced a restriction on the duration of each state in a phoneme, which must not be 0ms. Therefore, the speaking rates of the synthesized voices were not the same as the target rates which we indicated to the engine.

B. Results

Figure 3 shows the average recall rates in the second experiment. The results of T11-T14 and T21-T24 correspond to the iterations of the task set in the trial T1 and T2, respectively. Significant difference between T11 and T12 was observed . The F-test statistic value was $F(2, 30) = 30.93$ at the significance level of $p < 0.05$.

The results indicate that the learning effect is significant only in the early stage of the trials.

Figure 4 shows the average recall rates in the trial T1 and T2 in the second experiment. One can see that the recall rates of the Fast model surpasses that of the Normal model, regardless of the learning effect.

Figure 5 shows the evolution on the three weeks of experiments of the average recall rates of four subjects. One can see that the learning effect sustains for three weeks.

Though the number of subjects was limited, we investigated the individual variations of the learning effect. The variation of the individual recall rate is especially large in

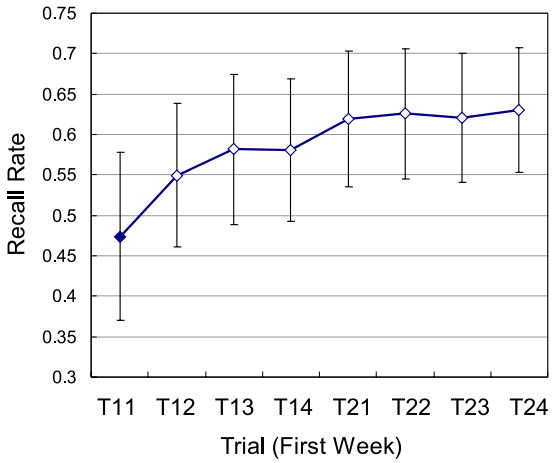

Fig. 3. Recall rates in the first week of the second experiment. The bars represent the distributions.

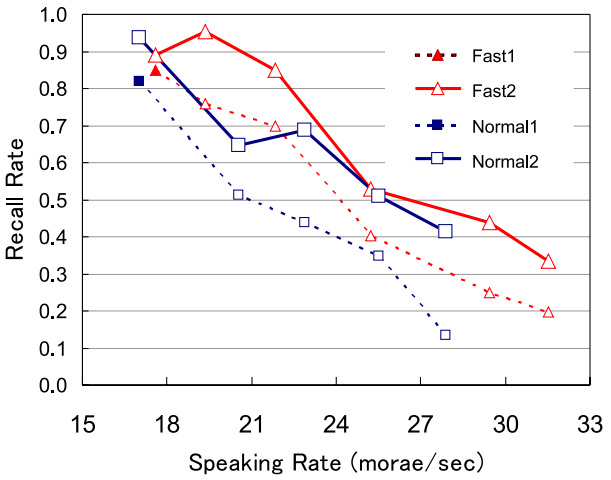

Fig. 4. Recall rates in trials T11 and T24 of the second experiment.

the sessions of the first two weeks. In the sessions of the third week, however, the variation decreases,as can be seen in Figure 5. The performance of a subject whose recall rate was lower than the other subjects in the first week came close to that of the others in the third week.

V. CONCLUSION

In this article, we investigated evaluation methods of Japanese TTS at fast speaking rates using a recall test of random numbers. We showed that the proposed method is effective for comparing the intelligibility of TTS.

Future work includes improvement of our HMM-based speech synthesizer and the rapidly-speaking statistical models, evaluation by subjects with visual disability, evaluation using a vocabulary other than the random numbers, and applications to voice interface systems/spoken dialog systems for the Web and home electric appliances [16].

VI. ACKNOWLEDGMENTS

This work was partially supported by the Ministry of Education, Culture, Sports, Science and Technology, Grant-in-Aid for Scientific Research on Priority Areas, #16091210, 2004.

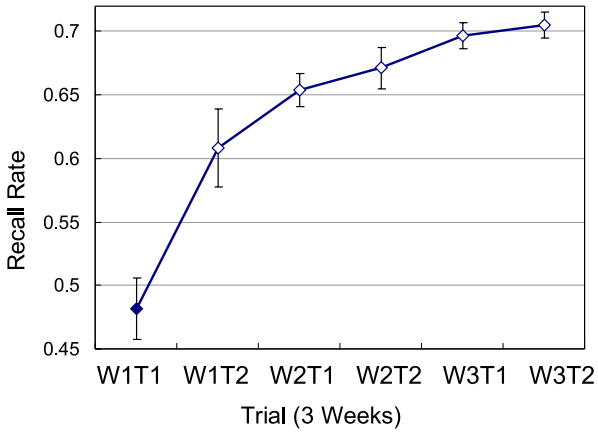

Fig. 5. Evolution of the recall rates during the three weeks of the second experiment. The bars represent the distributions.

REFERENCES

[1] T. Watanabe, "A Study on Voice Setting of Screen Readers for Visually-Impaired PC Users," *The IEICE Transactions on Information and Systems, Pt.1,* Vol.J88-D-I, No.8, pp.1257-1260, Aug 2005 (in Japanese).

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

[3] C. Asakawa, H. Takagi, S. Ino, T. Ifukube, "Maximum listening speeds for the blind," *Proceedings Conference of International Community for Auditory Display 2003,* pp. 276-279, 2003.

[4] C. Asakawa, H. Takagi, S. Ino, T. Ifukube, "The Optimal and Maximum Listening Rates in Presenting Speech Information to the Blind," *Journal of Human Interface Society,* Vol.7, No.1, pp.105–111, 2005 (in Japanese).

[5] http://hil.t.u-tokyo.ac.jp/∼galatea/

[6] http://www.astem.or.jp/istc/

[7] http://hts.ics.nitech.ac.jp/

[8] http://chasen.aist-nara.ac.jp/

[9] T. Nishimoto, S. Sako, S. Sagayama, K. Oda, T. Watanabe, "Evaluation of text-to-speech synthesizers at fast speaking rates," *IEICE Technical Report,* WIT2005-5, pp.23-28, May 2005 (in Japanese).

[10] T. Nishimoto, S. Sako, S. Sagayama, K. Ohshima, K. Oda, T. Watanabe, "Listener's familiarizations to text-to-speech synthesizers at fast speaking rates," *Proc. ASJ Autumn Conf.,* 3-6-14, pp.355-356, Sep 2005 (in Japanese).

[11] K. Ohshima, T. Nishimoto, T. Watanabe, "Relationship between oral comprehension and hearing experience of fast TTS for the visually disabled," *IEICE Technical Report,* WIT2005-43/SP2005-81, pp.19-24, Oct 2005 (in Japanese).

[12] S. Sako, T. Nishimoto, S. Sagayama, "A study on rapid speech synthesis using HMM-based speech synthesis technique," *Proc. ASJ Autumn Conf.,* 3-6-15, pp.357-358, Sep 2005 (in Japanese).

[13] G.E. Legge, C. Grosmann, C.M. Pieper. "Learning unfamiliar voices," Journal of Experimental Psychology: Learning, Memory and Cognition, 10, 298-303, 1984.

[14] T.J. Palmeri, S.D. Goldinger, D.B. Pisoni, "Episodic encoding of voice attributes and recognition memory for spoken words," *Journal of Experimental Psychology: Learning Memory and Cognition,* 19, pp.309–328, 1993.

[15] L.C. Nygaard, M.S. Sommers, D.B. Pisoni, "Speech perception as a talker-contingent process," Psychological Science, Vol. 5, No. 1, pp.42–46, Jan 1994.

[16] T. Watanabe, M. Yasumura, K. Oda, T. Nishimoto, "Basic research on speech recognition for the visually impaired and its application to voice interaction," *IEICE Technical Report,* WIT2004-74, pp.7-12, Mar 2005 (in Japanese).