The Effect of Learning on Listening to Ultra-Fast Speech

Takuya Nishimoto[a], Yukika Kariya[b] and Takayuki Watanabe[b]

[a]Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, 113-8656 Tokyo, Japan
[b]Department of Communication, Tokyo Woman's Christian University, 2-6-1 Zenpukuzi, Suginami-ku, 167-8585 Tokyo, Japan
nishi@hil.t.u-tokyo.ac.jp

We investigated the intelligibility of ultra-fast speech which may be used for screen reader for persons with visual disability. The subjects were 35 women who are university students and are not visually/hearing impaired. They were divided into four groups and they listened to 150 words with the speed of approximately 20 morae/sec. The vocabulary contained the tasks of high and low familiarity words, and the orders of tasks were different by the groups. Four morae Japanese words from the FW03 database were used as the vocabulary of the recall test. As a result, significant learning effect was observed in cases where the subject listened the high familiarity words in succession. This indicates that the learning effect to the ultra-fast speech is promoted when the mental lexical access is easy. We also investigated the mental workloads of the listening task using the NASA-TLX method. As the results, significantly high workload scores were observed at the listening of low familiarity words. The results also suggested that the mental workload decreases when the subject was aware that the mental lexical access was difficult.

1   Introduction

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. Screen-readers help the persons with visual disability to access the Web and read or write e-mails. 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.

Asakawa et al. [2, 3] 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.

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[4]. The results showed that the learning effect was significant in the early stage of the trials and the effect sustained for several weeks. We also investigated the effects of age-related effects on the same task[5]. Elderly persons can recall fast speech at some levels. However, their average recall rates are lower than the young university students and the individual differences are significant. The results of this tasks we consider are affected by the difficulty of auditory perception itself and the difficulty of recall the numbers in the correct order. Leaving out the effect of latter factor, it turned out that the task performances of elderly persons and young students are almost same. The learning effects of the elderly persons are not significant in either case, though those of the young students are significant for several weeks.

On the basis of past experiment, it is necessary to carry out experiments for more general vocabulary. This paper investigates the effects of word familiarity of the stimulus in this learning effect. We also examine the mental workload during listening. The process of learning of subject can be investigated in detail by measuring mental workload, because the subject may increases the mental workload during the active learning. Even if listener was able to hear the fast talking voice, it is undesirable for listener when the mental workload increases.

2   Mental Workload Measurement

Both physiological methods and objective evaluation techniques (e.g. dual task method) are well-known techniques to measure mental workload. In comparison with these techniques, we can easily carry out subjective assessment. Subject may be aware of difference of small workload that we cannot compare in objective evaluation. In such a case, more sensitive result is provided by subjective assessment. However, these results provided from introspectiveness of subject may not be reliable.

NASA-TLX[6, 7] is a subjective technique to evaluate mental workload with the following factors;

- Mental demand
- Physical demand
- Temporal pressure
- Effort
- Frustration level
- Performance

At first, the subject evaluates importance of each factor by doing the rankings. Then the subject accomplishes the work, followed by the evaluations of the workload for each factor. In our study, the subject inputs value from 0 to 100 by operation of scroll bar for each factor. Based on the ranking, we calculate weighted mean workload score (WWL). Magnitude of the most important factor and the most unimportant factor is 6 and 1 respectively.
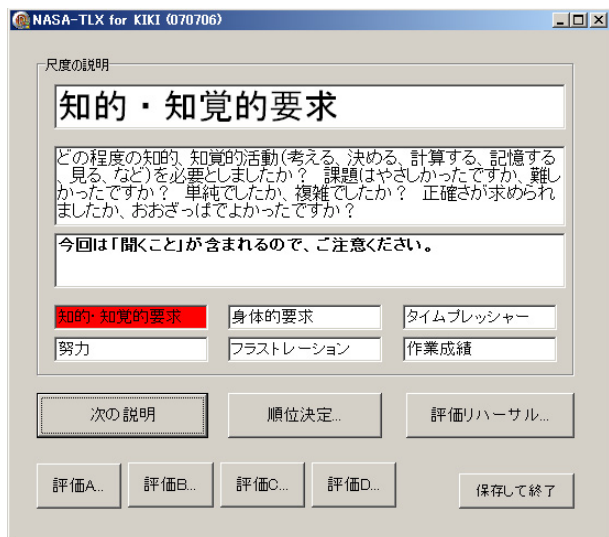
Figure 1: The NASA-TLX main menu with explanations of workload factors.

3 Familiarity-Controlled Words

FW03[8] is an audio data set for word intelligibility test. Vocabulary of this data is made based on Japanese word familiarity database. The word familiarity data were made by work to give value of familiarity with hands in 80,000 words of Japanese dictionary. The words of 4 morae which accent were controlled were chosen for every 4 groups corresponding to different familiarity. 20 groups of 50 words were made in consideration of balance of phoneme. This vocabulary was read aloud by for each 2 speaker of man and woman, and it was recorded. The speed to read is controlled in this recording. Calibration of level is accomplished for this audio data.

4 Preliminary Experiment

4.1 Experimental Setup

In the first experiment, we verify appropriateness of evaluating speech listening task with NASA-TLX. Furthermore, we verify that word familiarity affects both intelligibility and mental workload in this experiment.

We developed NASA-TLX software for Windows. Figure1 is screen of menu and explanation of each factor, and Figure2 is screen of evaluation of each factor. We designed it to make subject be conscious of magnitude of value between tasks. In other words, in evaluation of several tasks, we designed screen so that subject could refer to past evaluation value. In explanation of factor, we supplemented the following explanations:

- The highest performance means that you can listen to all words and write down correctly.

- Mental demand includes the workload to listen to sounds.

- Physical demand includes the workload to listen to sounds and to write characters.

We evaluated speech included in FW03 database. Conditions with regard to word familiarity are as follows.
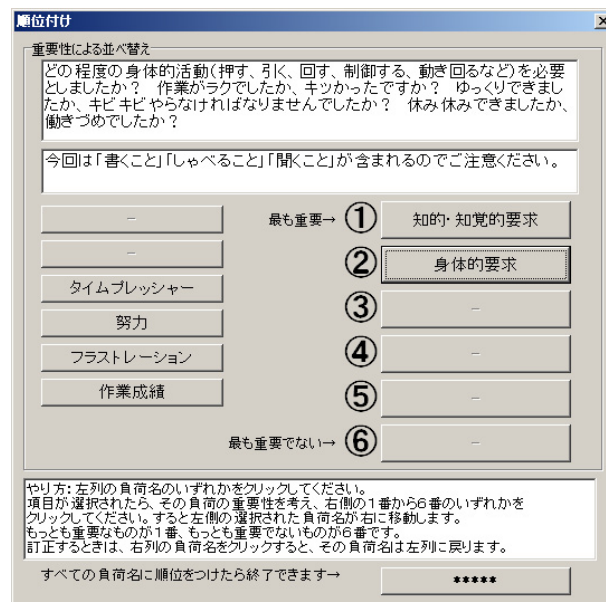

Figure 2: The evaluation of workload factors.

- FH : Word familiarity is from 7.0 to 5.5.

- FL : Word familiarity is from 2.5 to 1.0.

Conditions about speech rate are as follows:

- S1 : Speech files that are included in FW03.

- S2 : Speech files that we changed to double speed using speech editing software Adobe Audition 2.0.

The audio files are 48KHz sample rate, 16bit monaural speech files of 1 male speaker.

The task consists of FH-S1, FH-S2, FL-S1 and FL-S2. Each task consists of 50 words which are chosen in consideration of phoneme balance. All the words consist of 4 morae. Order of the words in each task is random. Average speed of S1 and S2 are 5.3 morae/sec and 10.8 morae/sec, respectively. The speech rate does not have significant difference between each group of FH and FL.

Subjects were 11 undergraduate students (women), and their native language was Japanese. They had neither visual impairment nor hearing impairment. They have not listened to the speech of the task.

We used a PC and a headphone for each subject. They listened to the speech played at intervals of 10 seconds and filled in answer sheets with the words that they heard. To avoid the effect of order, we divided the subjects to one group which listened to the fast speech earlier and another group which listened to the ordinary speech earlier.

4.2 Results

The result of 10 subjects were analyzed. A subject who did inappropriate answer was excluded. Figure 3 shows the distribution of word intelligibility and WWL in each task of the subjects. Figure 4 shows the distribution of word intelligibility and N-WWL (Normalized WWL). In this normalization, we transformed the average and standard deviation of WWL of each subject to 50 and 10, respectively. Table 1 shows Average(SD) of Intelligibility, WWL and Normalized WWL. The t-test (two
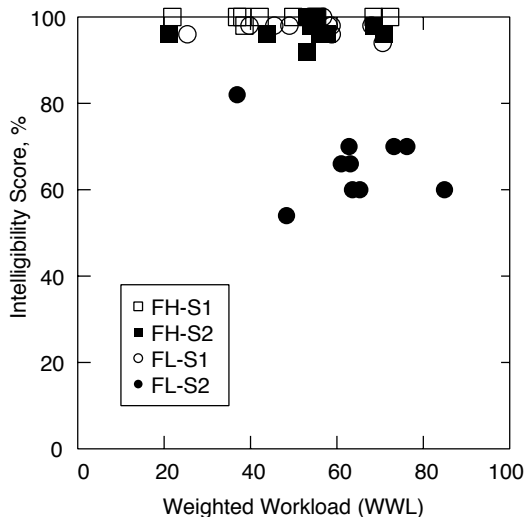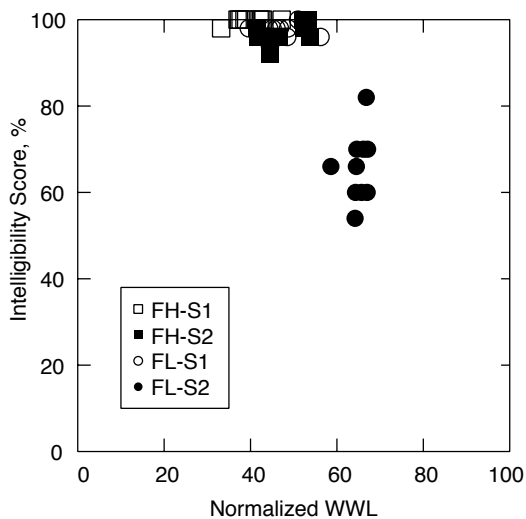
Figure 3: WWL and intelligibility score (Exp.1)

Figure 4: N-WWL and intelligibility score (Exp.1)

tailed distribution, level of significance = 5%) among the groups showed significant difference in all combinations of the groups, except the case between FH-S2 and FL-S1.

These results showed appropriateness of our experimental method. In addition, it was suggested that we could normalize individual variations using the adjusted standard deviation score of WWL.

5 Experiment on Learning Effects

5.1 Experimental Setup

In the second experiment, we examine relations of learning effect in listening of fast speech and word familiarity. The rate of speech which we used in this experiment is four times faster than the original speech. Condition about word familiarity is the same as experiment 1. In

Table 1: Average(SD) of Intelligibility, WWL and N-WWL (Exp.1)

Task	Intelligibility	WWL	N-WWL
FH-S1	99.8(0.6)	47.9(14.6)	39.7(3.7)
FH-S2	96.8(2.2)	53.3(13.0)	48.3(5.0)
FL-S1	97.4(1.6)	53.1(12.9)	47.1(4.2)
FL-S2	65.8(7.5)	63.5(12.9)	64.9(2.4)

Table 2: Experimental setup of Exp.2.

Group	Trial 1	Trial 2	Trial 3
L-L-L	FL-V1	FL-V2	FL-V3
H-H-L	FH-V4	FH-V5	FL-V3
L-L-H	FL-V1	FL-V2	FH-V6
H-H-H	FH-V4	VH-V5	FH-V6

other words we use FH and FL in this experiment. We prepared three sets of vocabulary at each level of word familiarity (i.e. vocaburary V1-V3 for familiarity FL, and V4-V6 for FH, respectively). The reason is we pay our attention to learning effect for voice, and to disturb learning for vocabulary.

The subjects were 35 undergraduate students (women). We divided them into 4 groups. Table 2 shows setup of the experiment. They listened to the speech played at intervals of 16 seconds. The subjects input words that they heard as alphabets (Roman characters) with keyboard of the PC.

5.2 Hypotheses

We wanted to verify the following hypotheses through this experiment:

1. Learning effect in this task would consist of (a) learning for task performance itself, (b) learning for acoustic stimulus as bottom-up information, (c) learning in regard to mental lexical access as top-down information.
   Under the condition that continues listening to words of low word familiarity, learning in regard to mental lexical access would be disturbed.
   In contrast, under the condition that continues listening to word of high word familiarity, learning of mental lexical access would be promoted. Therefore we would observe remarkable learning effect in HHH condition in comparison with LLL condition.

2. Under the condition that can utilize much top-down information, there becomes little mental workload. Subjects listening to speech of high word familiarity first notice that top-down information is available.
   Therefore, as for subjects listening to speech of low word familiarity after speech of high word familiarity, remarkable increase of mental workload will be seen.
   In contrast, subjects listening to speech of low word familiarity first will know that top-down information is not available. Therefore, as for them,

Table 3: Average (SD) of intelligibility, % (Exp.2)

Group	Trial 1	Trial 2	Trial 3
L-L-L	56.2 (8.4)	65.5 (6.1)	66.4 (6.8)
H-H-L	77.7 (5.5)	80.9 (3.1)	63.4 (7.3)
L-L-H	57.6 (6.5)	64.3 (5.4)	85.0 (5.8)
H-H-H	74.7 (8.8)	81.5 (6.3)	86.4 (5.2)

Table 4: Average (SD) of N-WWL (Exp.2)

Group	Trial 1	Trial 2	Trial 3
L-L-L	48.5 (11.2)	52.2 (7.9)	49.3 (10.2)
H-H-L	45.3 (9.2)	44.6 (6.6)	60.1 (4.2)
L-L-H	49.0 (6.0)	55.9 (8.6)	45.1 (11.4)
H-H-H	54.9 (10.8)	52.9 (7.7)	42.2 (5.6)

their mental workload will not reduce when they will listen to speech of high word familiarity later because we do not use top-down information intentionally.

5.3 Results

We counted the results of this experiment with respect to each morae which constituted the word. Table 3 and 4 show the average (SD) of intelligibility (%) and Normalized WWL (N-WWL), respectively. Figure 5, 6, 7 and 8 show the average intelligibility and N-WWL of each trial for the group L-L-L, H-H-L, L-L-H and H-H-H, respectively. The vertical and horizontal bars depict the standard deviations.

Results of t-test of two tailed distribution concerning word intelligibility are as follows:

- In result of group L - L - L, significant difference was seen only between T1–T2 (p=0.004).

- In case of H-H-H, significant differences were seen both between T1–T2 (p=0.015) and T2–T3 (p=0.014).

- In case of H-H-L, significant differences were seen both between T1–T2 (p=0.021) and T2–T3 (p=0.000). The latter difference can be assumed that the learning effect was canceled out by the effect of low word familiarity.

- In case of L-L-H, significant difference were seen both between T1–T2 (p=0.001) and T2–T3 (p=0.000). The latter difference can be assumed as the result of the learning effect and the effect of high word familiarity.

Results of t-test of N-WWL are as follows:

- In cases of L-L-L and L-L-H, no significant difference was observed.

- In case of H-H-H, however, significant difference was observed only between T2–T3 (p=0.010).

- In case of H-H-L, significant differences were also seen between T1–T3 (p=0.011) and T2–T3 (p=0.000), respectively.
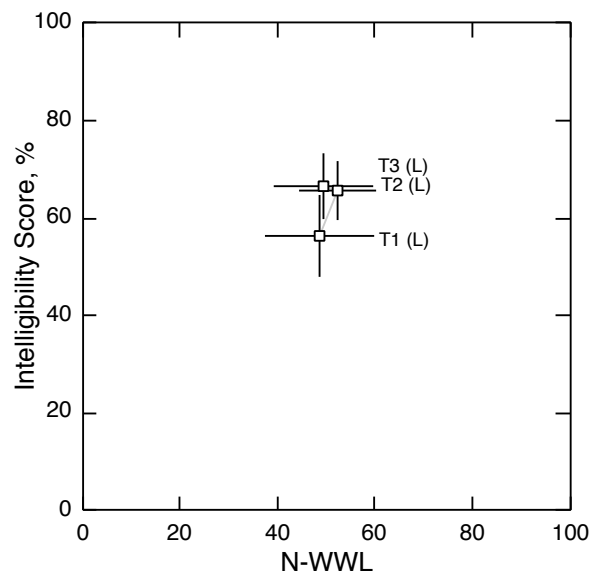

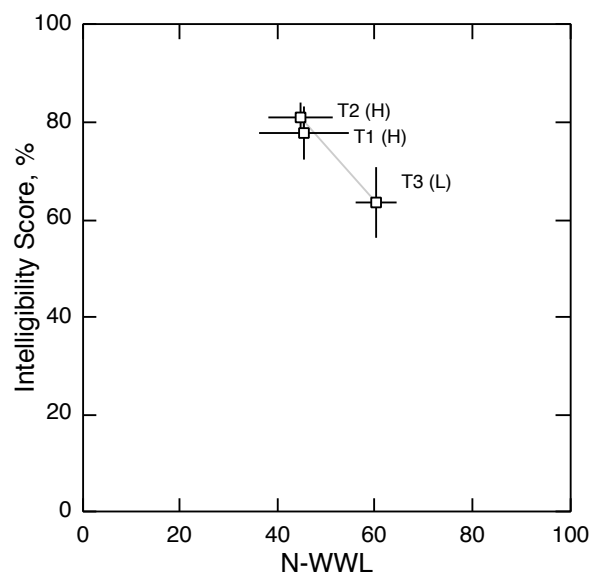Figure 5: N-WWL and intelligibility (L-L-L of Exp.2)

Figure 6: N-WWL and intelligibility (H-H-L of Exp.2)

These results support the hypothesis described in section 5.2. It cannot be said that we can listen to the speech of high word familiarity well after we get used to the speech of low word familiarity. We were not able to verify the opposite effect; i.e. the exercise with speech of high word familiarity would be effective for speech of low word familiarity.

6 Conclusion

We investigated learning effect in listening of fast speech with NASA-TLX. We investigated in particular effect of word familiarity using speech database FW03. Important findings that we obtained in this research are as follows. Particularly remarkable learning effect was provided by continuing listening to speech of high word familiarity. Mental workload was not affected by difficulty of task. Mental workload increased by carrying
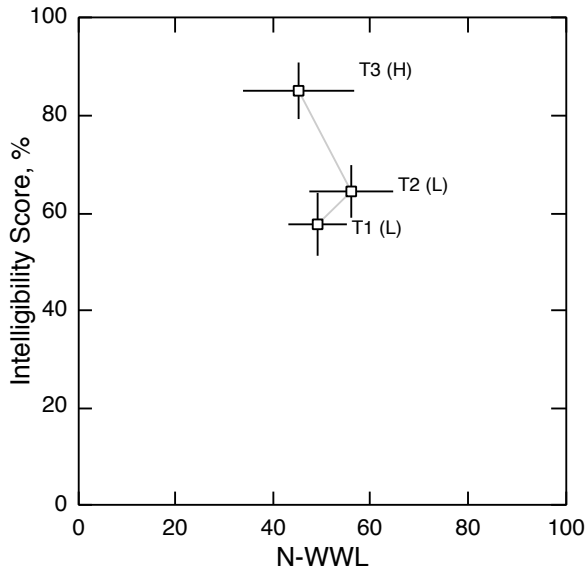
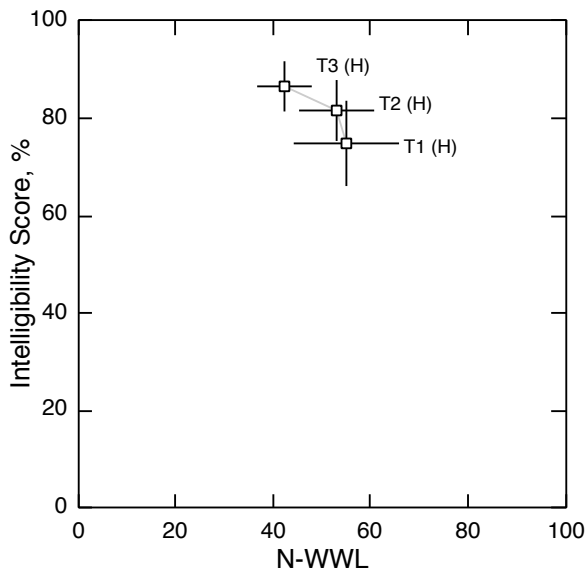Figure 7: N-WWL and intelligibility (L-L-H of Exp.2)

Figure 8: N-WWL and intelligibility (H-H-H of Exp.2)

out task with less top-down information.

Our future work includes the following topics:

- Examination of the effect of learning in longer period.

- Examination of difference of the effect by age and gender.

- Effects of smaller changes of speaking rate and word familiarity.

- Effects of instructions to subjects in the experiments.

- Applications to speech synthesis technology and language study education.

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] 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.

[3] 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).

[4] T. Nishimoto, S. Sako, S. Sagayama, K. Ohshima, K. Oda, T. Watanabe: "Effect of Learning on Listening to Ultra-Fast Synthesized Speech," Proceedings of the 28th IEEE Engineering in Medicine and Biology Society Annual International Conference (EMBC2006), pp.5691-5694, New York, Sep 2006.

[5] T. Nishimoto, T. Watanabe: "The Learning Effects and Age-Related Effects on Listening to Ultra-Fast Speech," Proceedings of Human Interface 2007, 3134, pp.937-942, Sep 2007 (in Japanese).

[6] S. G. Hart, L. E. Staveland: "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research," in P. A. Hancock and N. Meshkati (Eds.) *Human Mental Workload,* Amsterdam, North Holland Press (1998).

[7] S. Haga: The Theory and Measurement of Mental Workload, Japan Publication Service, 2001 (in Japanese).

[8] S. Amano, T. Kondo, S. Sakamoto, Y. Suzuki: Japanese Speech Dataset for Familiarity-Controlled Spoken-Word Intelligibility Test (FW03), NII Speech Resources Consortium (2006).