Investigation into the Effect of High-Speed Presentations of Educational Visual Content Utilizing Synthetic Speech

Toru Nagahama
Waseda University
tnagahama@aoni.waseda.jp

Masahiro Makino
Waseda University
ma19941002@asagi.waseda.jp

Yusuke Morita
Waseda University
ymorita@waseda.jp

Abstract: This study aims to clarify the effect of presenting educational visual content utilizing synthetic speech at a high speed. In the experiment, 40 university students were shown visual content dealing with declarative knowledge in 4 conditions: actual speed (1x) synthetic speech, double speed (2x) synthetic speech, actual speed (1x) normal speech, and double speed (2x) normal speech. An analysis of the comprehension test results showed no significant difference in the learning effect according to presentation condition, suggesting that speed and speech factors may have no impact on the learning effect. The results of a subjective questionnaire indicated that whereas the perception of normal speech as strange tends to be affected by the speed factor, the perception of synthetic speech as strange does not tend to be affected by that factor. An analysis of the interview results indicated that while learners found synthetic speech unnatural in terms of inflection and intonation when it was presented at actual speed, at high speed this unnatural impression was alleviated, and the speech became more acceptable to listeners.

Keywords: Educational visual content, Learning effect, Synthetic speech, High-speed presentation

1. INTRODUCTION

Massive open online courses (MOOCs) have gained global prominence in recent years (Waldrop 2013). Studies of the learning history of those taking MOOCs have shown that they spend longer viewing visual content than engaging in other learning activities related to MOOCs (Breslow et al. 2013, Kizilcec et al. 2013). In addition, Guo et al. (2014) used the learning history of MOOC students to show that the rate of studying visual content in which the lecturer's delivery is fast is higher than in the past. In addition, it was shown that the rate of studying was high for visual content lasting the relatively short time of 6–9 minutes.

Regarding the length of time spent viewing visual content, Aoyagi et al. (2005) showed that when studying themes that are relatively easy to understand, the same learning effect was obtained at an accelerated speed as at an ordinary speed. Also, Nagahama and Morita (2017) conducted experiments focusing on the variable speed function of MOOCs, in which visual content dealing with simple knowledge structures was presented at a fast speed. As a result, it was suggested that differences between presentation at actual speed and at double (2x) speed had no impact on the learning effect. This implies that with visual content dealing with learning themes that are comparatively easy to understand, the same learning effect is obtained from a high-speed presentation as from one at the original speed.

Attempts have been made to utilize synthetic speech when producing educational visual content (Kaburagi et al. 2003). Iwasaki and Ohashi (2015) prepared two versions of a narration to some visual content—one with the voice of the relevant lecturer (normal speech) and one with a synthetic voice—and conducted a comparative experiment. They found that evaluations of the synthetic speech were not positive because "it had no inflection and was monotonous" or the student "was distracted because the intonation and pronunciation felt strange."

Up to now, evaluations of speech simulation have been done from the viewpoint of understandability and naturalness (Watanabe 1989, Kasuya 1992). First, the understandability of synthetic speech was evaluated from the viewpoint of the degree of intelligibility at the level of phoneme, syllable, word, and sentence. It was reported that in highly understandable synthetic speech, the linguistic information was accurately conveyed (Pisoni et al. 1985, Higuchi et al. 1989, Watanabe 1989). The naturalness of synthetic speech can be evaluated from three viewpoints in its synthesis, namely i) segmental characteristics: adherence to the rules on pronouncing vowels and consonants including abnormal pronunciation such as devocalization, lengthening, nasalization, and omission of vowels; ii) prosodic characteristics: duration of morae and phonemes, accent, pauses, inflection, and loudness; and iii) voice

quality: smoothness, "noise," and overall impression. It was reported that the higher the perceived naturalness of synthetic speech, the closer it was to normal speech (Hieda 1988).

Kasuya et al. (1989) conducted interviews regarding the relationship between the understandability and naturalness of synthetic speech and its effectiveness in achieving a task. The interviews were with newspaper staff carrying out proofreading tasks using synthetic speech. Some participants expressed the opinion that "if the content is clear, when you get used to it, the unnaturalness of synthetic speech is no longer annoying."

Kumagami and Kasuya (1991) asked participants executing a light task to answer questions posed by a normal voice and two synthetic voices of differing naturalness. They found that on the first time hearing, the task was completed more slowly with the synthetic speech than with the normal speech, but that on the second and subsequent hearings, there was no significant difference in task speed between the synthetic speech and normal speech. In addition, they found no effect of differences in the naturalness of synthetic speech. These results suggest that the impact on task efficiency of inadequate naturalness in synthetic speech is smaller than the impact of understandability. In addition, it was suggested that an increase in the number of hearings allows acclimatization to (familiarity with) synthetic speech.

Meanwhile, there are existing studies that show that the most suitable synthetic speech presentation speed differs depending on the application. For example, Kasuya et al. (1991), in a proofreading task utilizing synthetic speech, asked participants to listen at 5 presentation speeds, from 340 to 680 morae per minute. They demonstrated that, when there was a high rate of inconsistency with the numbers read by the synthetic voice and the printed numbers, the high-speed speech production was poorly evaluated, and the slow-speech production was preferred. In addition, Shimahara (2000) sped up the presentation to visually impaired people unable to "speed read" of a synthetic voice at the part-of-speech level using syntactic information. He found that some users acquired the ability to "speed listen." In addition, Watanabe (2005) investigated the use of synthetic speech in screen readers for visually impaired people and found that many users set the presentation speed of their screen reader at the maximum (around 2x normal speed).

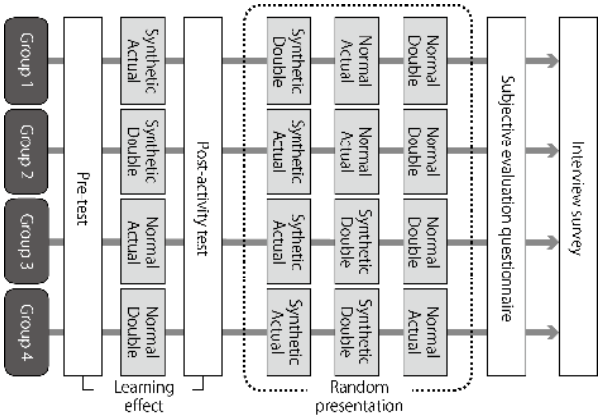

Figure 1. Experimental procedure.

However, so far there have been hardly any studies clarifying the effect of changing the presentation speed of educational visual content that uses synthetic speech. In addition, the usefulness of high-speed presentation of visual content has been verified in a number of studies (Nagahama and Morita, 2017). Thus, the aim of this study is to clarify the effect of high-speed presentation of visual content using synthetic speech.

2. METHOD

2.1. Overview of the experiment

Visual content was presented in four conditions: with synthetic speech at actual speed (1x), with synthetic speech at double speed (2x), with normal speech at actual speed (1x), and with normal speech at double speed (2x). The participants in the experiment were 40 students (24 male, 6 female; average age 21.4 [SD=0.9]) at a private university in the Tokyo area. Please note that to take account of the impact of familiarity with synthetic speech, we confirmed (self-certification) with all participants that they had no prior experience of learning to utilize synthetic speech and did not utilize synthetic speech in their daily lives.

Figure 1 shows the experimental procedure. First, participants were sorted into four groups of equal size. Next, a pre-test was carried out to confirm the existing level of knowledge before the learning activity. Next, visual content in each of the four conditions was shown to each group (Table 1). They were asked not to pause or rewind the content. Next, a post-activity test with the same content as the pre-test was carried out to measure the effect on learning. Thereafter, participants were randomly shown visual content of the various conditions so that they all viewed all four presentation conditions and then answered a subjective evaluation questionnaire. Finally, an interview survey was carried out.

2.2. Overview of experimental videos

The visual content presented in the normal speech condition (normal speech videos) was the same as that used by Nagahama and Morita (2017). Its subject matter was network structure as taught in information studies at high school, and the lecturer was a currently practicing high school teacher of information studies at a private high school in Chiba Prefecture. In addition, with reference to Fukumori (2008), we measured the speed of the speech in morae (a mora is a sound segment unit in phonics with a certain

Table 1. Initial lecture video presentation conditions by group.

Group		Age	Male	Female	Initial lecture video presentation condition	
					Speed	Speech
Group 1	(*n*=10)	21.6 (0.7)	5	5	Actual	Synthetic
Group 2	(*n*=10)	21.3 (0.7)	7	3	Double	Synthetic
Group 3	(*n*=10)	21.6 (1.3)	6	4	Actual	Normal
Group 4	(*n*=10)	21.1 (0.9)	6	4	Double	Normal

Table 2. Normal speech and synthetic speech overview.

Slide 【subject matter】	Slide presentation time (seconds)	No. of morae		Speech presentation time (seconds)	
		Synthetic speech	Normal speech	Synthetic speech	Normal speech
Slide 1 【Networks】	89	455	466	60	60
Slide 2 【Components】	53	255	262	34	34
Slide 3 【Protocol】	140	781	792	104	98
Slide 4 【IP address】	112	616	631	83	77
Slide 5 【DNS server】	41	225	230	30	27
Slide 6 【URL creation】	79	411	428	54	54
Average (*SD*)	85.7 (33.6)	457.2 (194.5)	468.2 (196.7)	60.8 (26.1)	58.3 (24.2)

temporal length) and found that it was 327.9 morae per minute.

The visual content presented in the synthetic speech condition (synthetic speech videos) was produced based on that used by Nagahama and Morita (2017). For the synthetic speech, we utilized the text-to-speech function of an iMac (Retina 5K, 27 inch) produced by Apple to transfer data to speech. The script to be read was produced from teaching materials in a normal speech video, but without slips of the tongue, auxiliary words, or fillers. In addition, the reading speed and reading voice were set using the default settings on the iMac. The voice data were edited to the match the voice production timing in the normal speech videos using Final Cut Pro X by Apple.

The presentation time of the normal speech and synthetic speech videos was 9 minutes and 12 seconds in the actual speed conditions and 4 minutes and 42 seconds in the double-speed conditions. In addition, six slides were created, excluding the introductory section. Table 2 summarizes the normal and synthetic speech information. The number of morae was higher for the synthetic speech than for the normal speech because fillers, auxiliary words, and slips of the tongue were excluded from the normal speech video's teaching content. In addition, the length of time that speech was presented (speech presentation time) was measured with a stop watch for both synthetic and normal speech; it was revealed that the synthetic speech presentation time was longer than the normal speech. Please note that the normal speech presentation time was measured with fillers and slips of the tongue in the teaching content excluded. Particles at the end of sentences were excluded from the measurement because the corresponding audio was too short.

2.3. Comprehesion test

We used Nagahama and Morita's (2017) comprehension test to measure the learning effect, administering a pre-test and a post-activity test. The comprehension test consisted of 20 questions (11 recall questions and 9 applied knowledge questions). One mark was awarded for each correct answer, with 20 marks being the maximum score. The recall questions were presented in the form of an information recall test and were intended to measure the volume of information retained by the participants after viewing the visual content. The applied knowledge questions included one multiple choice question, five recognition questions, and three true or false questions, and were intended to measure the ability to apply knowledge learned from the visual content to new problems.

Table 4. Average increase in score on the comprehension test.

	Synthetic speech		Normal speech		*F* value		
	Actual	Double	Actual	Double	Speech	Speed	Interaction
Total score	8.3 (2.2)	6.2 (3.0)	7.7 (2.2)	7.1 (2.3)	0.0 *ns*	2.7 *ns*	0.8 *ns*
Recall score	6.0 (2.1)	4.9 (1.8)	5.0 (1.7)	4.7 (1.9)	0.9 *ns*	1.3 *ns*	0.4 *ns*
Applied knowledge score	3.0 (1.3)	1.8 (1.0)	4.4 (0.9)	2.0 (1.2)	15.8 *ns*	137.2 *ns*	13.5 *ns*

**: $p < .01$, *: $p < .05$, +: $p < .10$

2.4. Subjective evaluation questionnaire

A paper question sheet was used in the subjective evaluation questionnaire. There were 25 items, consisting of 1 item in Nagahama and Morita's (2017) "comprehension" category, 3 in their "lecturer" category, 3 in their "concentration" category, 1 in their "audibility" category, 2 in their "viewability" category, 2 in their "presentation speed, time taste" category, and 3 in their "content taste" category, to give a total of 15 items, plus 10 items added for this study.

The questions were answered on a five-point scale, with five meaning "strongly agree," four meaning "somewhat agree," three meaning "neither agree nor disagree," two meaning "somewhat disagree," and one meaning "strongly disagree."

2.5. Interview survey

The interview survey was carried out using the semi-structured method, with interviews lasting 5–10 minutes and being conducted by the authors. The questions related to the differences between normal speech and synthetic speech in the actual speed condition and the double speed condition.

3. RESULTS AND DISCUSSION

3.1. Confirmation of homogeneity

To confirm homogeneity between the four groups, we carried out a one-way ANOVA on the scores in the pre-test. This showed no significant difference between the groups (F (3, 36) = 0.61, *n.s.*). This confirmed that the existing knowledge of the teaching content prior to the learning activity was at the same level in all four groups.

3.2. Comprehensiton test analysis

We collected the overall scores on the comprehension test, the scores for the recall questions (recall score), and the scores for the applied knowledge questions (applied knowledge score), which are shown in Figure 2. We conducted a two-way ANOVA (Table 4) regarding the average rise in overall score, recall score, and applied knowledge score on the comprehension test, using the speech factor (relating to type of speech in the visual content) and the speed factor (relating to the speed of the presentation of the visual content).

First, no interaction was visible regarding the growth in the overall score (F (1, 36) = 0.84, *n.s.*). When we tested the main effect, we found no significant difference for the speech factor (F (1, 36) = 0.03, *n.s.*), and no significant difference for the speed factor (F (1, 36) = 2.73, *n.s.*). This clarifies that the speech and speed factors did not influence the rise in overall score.

Next, no interaction was visible regarding growth in the recall question score (F (1, 36) = 0.41, *n.s.*). When we tested the main effect, we found no significant difference for the speech factor (F (1, 36) = 0.92, *n.s.*), and no significant difference for the speed factor (F (1, 36) = 1.25, *n.s.*). This clarifies that the speech and speed factors did not influence the rise in recall score.

Next, no interaction was visible regarding the growth in applied knowledge score (F (1, 36) = 0.32, *n.s.*). When we tested the major effect, we found no significant difference for the speech factor (F (1, 36) = 1.48, *n.s.*), and no significant difference for the speed factor (F (1, 36) = 1.11, *n.s.*). This shows that the speech and speed factors did not influence the rise in applied knowledge score.

In the overall, recall, and applied knowledge scores, there was no significant interaction between the speech and speed factors and no significant main effect. This suggests that under the conditions in this experiment, the speech and speed factors had no influence on the learning effect.

3.3. Analysis of subjective evaluation questionnaire

Relating to the answers in the subjective evaluation questionnaire, we collected scores for each condition and calculated the average value for each item. To investigate the factorial structure, we conducted an exploratory factor analysis (maximum likelihood method, promax rotation). As a result, we identified four factors from the sharp decline in the scree plot. These factors having been indicated, we carried out a factor analysis (excluding items with a loading of less than 0.35) and obtained 4 factors and 21 items.

Table 5 shows the factor analysis results. Factor 1 was comprised of items relating to the usefulness of the visual content (such as Item 5 "I was able to

Table 5. Factor analysis table.

			Factor 1	Factor 2	Factor 3	Factor 4
Usefulness of visual content (α= .86)	Q5	I was able to concentrate and listen	**.85**	.25	.08	-.15
	Q1	I understood the teaching content	**.84**	-.01	.11	-.34
	Q2	The explanation was clear	**.83**	-.17	.17	.07
	Q4	The explanation was well structured	**.66**	-.22	-.03	.11
	Q3	The lecturer's voice was intelligible	**.65**	-.09	-.02	.17
	Q21	The voice was easy on the ears	**.55**	.04	-.17	.13
	Q12	I would like to use this presentation condition to study again	**.52**	.13	.02	.19
	Q11	There were places where I wanted a slower explanation	**- .39**	.25	.03	.19
Perceived strangeness of presentation speech (α= .77)	Q22	The voice inflection was annoying	.03	**.98**	-.05	-.11
	Q23	The voice production intonation was annoying	.02	**.90**	.04	-.12
	Q20	There was a high volume of spoken information	.01	**.41**	.03	.20
Presentation information burden (α= .72)	Q7	The screen flicker was annoying	.20	.11	**.83**	-.07
	Q14	There were a lot of charts and tables on the slides	-.13	-.00	**.62**	.23
	Q6	My eyes became tired while watching	.01	-.01	**.58**	.03
	Q18	The presentation speed was appropriate	-.09	.13	**- .49**	-.44
	Q25	The timing of the voice production was appropriate	.34	-.09	**- .42**	.08
Form and understandability of the visual content (α= .66)	Q16	I liked being able to see the captions	.05	.10	-.01	**.70**
	Q24	The gaps between the voice production were unnatural	-.16	.17	.23	**- .49**
	Q19	The spoken information deepened my understanding	.29	.29	-.30	**.48**
	Q10	I followed the textual information with my eyes with difficulty	-.01	.43	.26	**.47**
	Q9	I focused on the visual information while viewing	.20	.16	-.13	**- .43**
	Q13	The text volume on the slides was low	.11	-.04	.32	**.42**
		Factor 1	—	-.44	-.25	.20
		Factor 2		—	.31	-.02
		Factor 3			—	.02
		Factor 4				—

concentrate and listen," Item 1 "I understood the teaching content," Item 3 "The lecturer's voice was intelligible," and Item 21 "The voice was easy on the ears"), and so it was dubbed the "usefulness of visual content" factor. Factor 2 was comprised of items relating to perceptions of strangeness in the presentation speech of the visual content (such as Item 22 "The voice inflection was annoying" and Item 23 "The voice production intonation was annoying"), and so it was dubbed the "perceived strangeness of presentation speech" factor.

Factor 3 was comprised of items relating to the burden of the presentation's visual content (such as Item 7 "The screen flicker was annoying," Item 14 "There were a lot of charts and tables on the slides," and Item 6 "My eyes became tired while watching"), and so it was dubbed the "presentation information burden" factor. Factor 4 was comprised of items relating to the form of the visual content and its understandability (such as Item 16 "I liked being able to see the captions," Item 24 "The gaps between the voice production were unnatural," and Item 10 "I followed the textual information with my eyes with difficulty"), and so it was dubbed the "form and understandability of the visual content" factor.

We investigated the reliability of the criteria using α coefficients; that for "usefulness of visual content" was .86, that for the "perceived strangeness of presentation speech" was .77, that for "presentation information burden" was .72, and that for "form and understandability of the visual content" was .66. Please note that when investigating the reliability of the criteria using α coefficients, items where the loading showed a minus value were inverted for processing.

3.3.1 Analysis of sub-scale scores for each factor

We calculated the average value of the items in each factor for each presentation condition and made them the respective sub-scale's score. We also carried out a two-way ANOVA for the speech and speed factors (Table 6).

Table 6: Subscale score for each factor

	Synthetic speech		Normal speech		*F* value		
	Actual	Double	Actual	Double	Speech	Speed	Interaction
Usefulness of visual content	3.1 (0.7)	1.8 (0.8)	4.2 (0.7)	2.2 (1.0)	32.3 **	170.2 **	8.5 **
Perceived strangeness of presentation speech	3.6 (0.9)	3.7 (1.1)	2.0 (0.8)	2.7 (0.9)	51.9 **	8.4 **	14.5 **
Presentation information burden	2.4 (0.6)	2.8 (0.7)	1.9 (0.6)	2.7 (0.7)	34.1 **	34.0 **	6.1 *
Form and understandability of the visual content	2.6 (0.7)	2.5 (0.6)	2.8 (0.6)	2.4 (0.6)	3.6 +	30.7 **	11.8 **

**: $p < .01$, *: $p < .05$, +: $p < .10$

First, interaction was seen as relating to the sub-scale score for the "usefulness of visual content" factor (F (1, 39) = 8.5, $p < .01$). An investigation of the simple main effect revealed a significant difference in the actual speed condition (F (1, 39) = 34.0, $p < .01$) and the double-speed condition (F (1, 39) = 5.5, $p < .05$), regarding the speech factor. Meanwhile, regarding the speed factor, a significant difference was seen for the normal speech condition (F (1, 39) = 103.7, $p < .01$) and the synthetic speech condition (F (1, 39) = 79.8, $p < .01$). These results showed that the degree of usefulness of the visual content as felt by the participants differed according to the speech and speed factors.

Next, an interaction was seen relating to the sub-scale score of the "perceived strangeness of presentation speech" factor (F (1, 39) = 14.5, $p < .01$). An investigation of the simple main effect revealed a significant difference in the actual speed condition (F (1, 39) = 54.3, $p < .01$) and the double-speed condition (F (1, 39) = 32.4, $p < .05$), regarding the speech factor. Meanwhile, regarding the speed factor, a significant difference was seen for the normal speech condition (F (1, 39) = 16.5, $p<.01$) and the synthetic speech condition (F (1, 39) = 0.3, *n.s.*). These results showed that while the degree of "perceived strangeness of the presentation speech" as felt by the participants differed according to the speed factor in the normal speech condition, in the synthetic speech condition, it was the same regardless of the speed factor.

Following on from this, an interaction was seen relating to the sub-scale score of the "presentation information burden" factor (F (1, 39) = 6.1, $p < .05$).

An investigation of the simple main effect revealed a significant difference in the actual speed condition (F (1, 39) = 30.9, p < .01) and the double-speed condition (F (1, 39) = 6.6, p < .05), regarding the speech factor. Meanwhile, regarding the speed factor, a significant difference was seen for the normal speech condition (F (1, 39) = 34.2, p < .01), and the synthetic speech condition (F (1, 39) = 20.2, p < .01). These results showed that the degree of "presentation information burden" felt by the participants differed according to the speech and speed factors.

Following on from that, an interaction was seen relating to the sub-scale score of the "form and understandability of the visual content" factor (F (1, 39) = 11.8, $p < .01$). An investigation of the simple main effect revealed a significant difference in the actual speed condition (F (1, 39) = 14.1, $p < .01$) and the double-speed condition (F (1, 39) = 0.4, *n.s.*), regarding the speech factor. Meanwhile, regarding the speed factor, a significant difference was seen for the normal speech condition (F (1, 39) = 38.5, $p < .01$) and the synthetic speech condition (F (1, 39) = 3.1, $p < .10$). These results showed that while subjective evaluations of the "form and understandability of the visual content" differed according to the speech factor in the actual speed condition, they were the same regardless of the speech condition in the double-speed condition.

3.4. Results of the interview survey

When all of the comments obtained via the interviews were summarized and counted, there were 196 in total. All of the comments obtained were classified into the four categories of "comprehension," "intelligibility," "acceptability to the listener," and "concentration," according to their key words. In doing this, with reference to Kasuya (1992), we classified comments relating to "comprehension" in the "intelligibility" category and those relating to "naturalness" in the "acceptability to the listener" category. As a result, 17 comments were classified under "comprehension," 51 under "intelligibility," 80 under "acceptability to the listener," and 32 under "concentration." Please note that 16 comments did not fit into any of the 4 categories and were classified under "other."

3.5. Disucussion of the results of the interview survey

The following discussion of the results of the interview survey is conducted from the two viewpoints of ease of listening and problems related to speeding up the speech.

3.5.1 Ease of listening to the presentation speech

Among the comments relating to the ease of listening to the presentation speech, there were positive comments regarding normal speech at actual speed relating to the presence of inflection, a sense of friendliness, and the presence of familiarity. There were also negative comments relating to being bothered by slips of the tongue and fillers.

Meanwhile, there were negative comments regarding normal speech at double speed, to the effect that “there were too many fillers” and “sped-up fillers are jarring to hear.” These comments point to the possibility, relating to the ease of listening, that the participants were positive about normal speech at actual speed, but fillers increased its unpleasantness when it was sped up.

Elsewhere, there were positive comments regarding synthetic speech at actual speed relating to the absence of slips of the tongue and fillers. There were also negative comments relating to the absence of inflection, the unnatural intonation, and the monotonous rhythm. However, although there were negative comments relating to the double-speed synthetic speech to the effect that the monotonous or high tone was jarring, there were also positive comments to the effect that “the aspects of synthetic speech that are strange at actual speed are no longer annoying” and “the unnaturalness of the voice is less annoying than at actual speed and it becomes easier on the ear.”

These comments suggest that when it comes to the ease of listening to the presentation speech, the participants found the synthetic speech unnatural in terms of inflection, intonation, and rhythm when the presentation was at actual speed, but that the sense of strangeness and unnaturalness was alleviated when the presentation was sped up and it became more acceptable to the listener.

3.5.2 Problems with the high-speed speech presentation

There were indications that any sense of strangeness and unnaturalness relating to the synthetic speech was alleviated by the high-speed presentation. However, there were comments relating to the double-speed synthetic speech to the effect that “it was fast, and I could pick out hardly anything” or “it was fast and very difficult to follow.” There were also comments along the lines of “I did absorb the information, but it was tiring” or “It was too fast and I gradually lost heart.” These comments imply that listening took a toll on the participants. Notably, comments to the effect that “I felt that the volume of spoken information was higher than the volume of information on the slides,” “I relied almost entirely on the slides,” and “I did not rely on the speech” suggest that processing auditory information took a particularly high toll on the participants.

Meanwhile, there were similar comments relating to the auditory burden of double-speed normal speech. In addition, some participants commented that they felt little difference between double-speed (2x) normal speech and double-speed (2x) synthetic speech. These comments imply that when presented at a high speed, normal speech takes on some of the qualities of synthetic speech.

The above discussion suggests the necessity of a thorough investigation into how to deal with the associated auditory burden before synthetic speech or normal speech are presented at a high speed.

4. CONCLUSION

You can use this document by merely doing the copy and paste your text over in this template. To read this template correctly, select "Print Layout" 3.under the "View" menu to show the two-column format.

The aim of this study was to clarify the effect of a high-speed presentation of educational visual content using synthetic speech. In the experiment, 40 university students were presented with visual content dealing with declarative knowledge in 4 conditions (actual speed synthetic speech, double-speed synthetic speech, actual speed normal speech, and double-speed normal speech).

An analysis of the comprehension test results suggested that neither the factor relating to the speech (speech factor) nor the factor relating to the presentation speed (speed factor) had any impact on the learning effect.

An analysis of the subjective evaluation questionnaire suggested that the subjective evaluations relating to the “usefulness of the educational visual content” and “the burden from the presentation information” differed according to the speech and speed factors. In addition, it was suggested that the subjective evaluations relating to the “perceived strangeness of the presentation speech” were not impacted by the speed factor in the synthetic speech condition. Furthermore, it was found that the subjective evaluations relating to the “form and understandability of the visual content” were not affected by the speech factor.

The interview survey results implied, with regard to the presentation’s ease of listening, that the learners found the synthetic voice unnatural in terms of inflection, intonation, and rhythm when hearing it at actual speed, but that when it was presented to them at high speed, this perceived unnaturalness was alleviated and it became more acceptable to the listener. In addition, it seems that the absence of fillers and slips of the tongue as found in normal speech increased the ease of listening of the synthetic speech.

From the above, it is clear that under the conditions in our experiment, when visual content using synthetic speech was presented at double-speed,

the same learning effect was achieved as with actual speed. It also shows that the unnaturalness relating to the inflection, intonation, and rhythm of synthetic speech as alleviated by speeding up the presentation speed, and there was an improvement in the acceptability to the listener and the intelligibility.

However, there was no thorough investigation of the naturalness of the inflection, pauses, and intonation of the synthetic speech used in this experiment. This was because the text-to-speech software used was one that is installed on a widely available computer, and so the assumption was made that intelligibility would be sufficiently guaranteed. In addition, Kumagami and Kasuya (1991) indicated that users rapidly get used to synthetic speech. In this study, we only asked the participants to listen to synthetic speech in each condition once, and we did not get to the stage of thoroughly investigating the impact of increased familiarity with synthetic speech.

There were thus limitations to the evaluation in this study. Therefore, a more comprehensive discussion is required regarding the effect of high-speed presentations of visual content utilizing synthetic speech; this discussion should encompass the various elements relating to synthetic speech.

REFERENCES

Aoyagi, S., Sato, K., Takada, T., Sugawara, T., & Onai, R. (2005). Evaluation of video skimming method to educational purpose movies. *Journal of Information Processing, 46(5)*, 1927-1305.

Breslow, L., Pritchard, D. E., Deboer, J., Stump, G. S. S., Ho, A. D., & Seaton, D. T. (2013). Studying learning in the worldwide classroom: Research into edX's first MOOC. *Research and Practice in Assessment*, *8*, 13-25.

Guo, P. J., Kim, J., & Rubin, R. (2014). How video production affects student engagement: An empirical study of MOOC videos. *In Proceedings of the First ACM conference on Learning*, 41-50.

Hieda, I. (1988). Subjective indices for evauation of synthesized voice. *Japan Ergonomics Society Research Journal*, *24*, 387-394.

Higuchi, N., Yamamoto, S., & Shimizu, T. (1989) Evaluation of intelligibility and naturalness of the synthetic speech generated with a Japanese speech synthesizer by rule. *Journal of the Institute of Electronics, Information and Communication Engineers*, *D-Ⅱ (J72-D-H)*, 1133-1140.

Iwazaki, K. & Ohashi, A. (2015). Active learning experiences in the fipped classroom. *Computer & Education*, *39*, 98-103.

Kaburagi, M., Uehashi, J., Asase, J., Kato, M., & Kang, M. (2003). Development of supporting system with speech engine for material creation and learning. *Japan Journal of Educational Technology*, *27(Suppl.)*, 141-144.

Kasuya, H. (1992). Assessment of speech synthesis technology. *The Journal of the Acoustical Society of Japan*, *48(1)*, 46-51

Kasuya, H., & Morita, K. (1991). Role of the speaking rate of synthetic speech produced by rule as an aid for proofreading. *The Journal of the Acoustical Society of Japan*, *47*, 96-98

Kasuya, H., Morita, K., & Kumagami, K. (1989). Investigation relating to evaluation of the quality of synthetic speech. *Report of a study funded by the Kakenhi Grant-in-Aid for Scientific Research in program (important areas, speech and language)*, PASL01-8-2.

Kizilcec, R. F., Piech, C., & Schneider, E. (2013). Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses. *In Proceedings of the Third International Conference on Learning Analytics and Knowledge, ACM*, 170-179.

Kumagami, K., Kasuya, H. (1991). Objective evaluation of user's adaptation process to synthetic speech produced by rule. *The Journal of the Acoustical Society of Japan*, *47*, 243-249.

Nagahama, T., & Morita, Y., (2017). An analysis of the effects of learning with high-speed visual contents. *Japan Journal of Educational Technology, 40(4)*, 291-300.

Nagahama, T., & Morita, Y. (2017). Effect analysis of playback speed for lecture video including instructor images. *International Journal for Educational Media and Technology*, *11(1)*, 50-58.

Pisoni, D. B., Nusbaum, H. C., & Green, B. G. (1985). "Perception of synthetic speech generated by rule." *In Proceedings of the IEEE 73*: 1665-1676.

Shimahara, S. (2000). 'Speed listening' – a fast reading system for visually impaired people using syntactic information. *Institute of Electronics, Information and Communication Engineers Technical Report, Fifth Well-being Information Technology Conference WIT00-28.*

Waldrop, M. M. (2013). Online learning: Campus 2.0. *Nature*, *495*, 160-163. http://dx.doi.org/10.1038/495169a (accessed 3.10.2018).

Watanabe, T.(2005). A study on voice settings of screen readers for visually-impaired PC users. *The IEICE Transactions on Information Systems Pt. 1, 88(8)*, 1257-1260, 2005-08-01.

Watanabe, T. (1989). Investigation relating to methods of evaluating rule-based synthetic speech using degree of work comprehension. *Institute of Electronics, Information and Communication Engineers Research Journal* , *J72-A*, 1503-1509