JACET Journal 53 (2011) 1–17

Comprehensibility and Naturalness of Text-To-Speech Synthetic Materials for EFL Listeners

HIRAI, Akiyo
University of Tsukuba
O'KI, Toshihide
Hakuoh University

Abstract

Text-to-speech (TTS) synthesis technology, which has rapidly advanced in recent years, can be an extremely useful tool to help EFL teachers make listening materials for their students. However, little has been researched regarding the use of synthetic speech created by a recent TTS program (e.g., *Globalvoice English*) as an alternative to human speech. Thus, this paper examines the quality of synthetic speech by comparing it with natural human speech. The comprehensibility and naturalness of the two speech types were measured by administering a listening multiple-choice test and a questionnaire. Results suggested that (1) even though natural speech tended to be comprehended better than synthetic speech, prior exposure to synthetic speech seemed to affect its comprehensibility, (2) synthetic speech was perceived to be almost as natural as human speech by both upper and lower proficiency groups, and (3) when compared with higher level students, more students in the lower level group tended to prefer synthetic speech. These results support the application of synthetic speech in creating materials for English education, though further studies are still necessary to apply the findings to other pedagogical and research contexts.

Key Words: text-to-speech synthesis, listening tests, comprehensibility, naturalness

Introduction

No one would disagree that life has become more convenient with the development of technology. One that has flourished in recent years and might interest English teachers is text-to-speech (TTS) synthesis. TTS synthesis is a computerized system in which written texts are synthetically transformed into speech. Since the system has made it extremely easy for users to create intended speech and edit on the computer, it has potential to reduce arduous procedures involved in preparing listening materials. This paper, thus, examines to what extent synthetic speech created by a TTS synthesis program (*Globalvoice English ver.2*) can be applicable to creating classroom listening material when compared to human voiced speech.

Advantages of Using TTS Synthesis Technology

EFL teachers and researchers sometimes need to create listening materials for teaching or assessing learners' aural/oral skills. For example, in the Story Retelling Speaking Test (SRST;

Hirai & Koizumi, 2009), in which the examinees retell a story after reading or listening to it, many short stories are needed to assess the examinees' proficiency levels. It is easy to find or edit written stories, but it is difficult to find and edit stories in the speech media. Another example can be found in studies that have investigated word recognition in a shadowing task, in which learners repeat an aural passage as soon as they hear it (e.g., Oki, 2010a, 2010b). In these studies, a relatively small number of semantically anomalous speeches are recorded by a human voice. However, a TTS synthesis program may make it much easier to not only produce more material but also control the phonological aspects of a speech, such as speed, volume, rhythm, and intonation. Controlling these aspects can produce highly reliable data.

As the examples mentioned above indicate, the TTS synthesis system can aid EFL teachers and researchers in creating listening tasks, depending on the aspect they wish to focus on.

Evaluation of Synthetic Speech Quality

A common question may be whether the TTS technology can ever produce quality speech that can replace human voiced speech. Azuma (2010) provides some evidence supporting the effectiveness of the technology. In his study, he presented 27 participants with human speech and five kinds of synthetic speech and let them select the speech that sounded the closest to one voiced by a human. As many as 25 participants chose either of the five kinds of synthetic speech. Such a high rate of success may be attributed to the technique called "concatenative approach" used in TTS synthesis programs, in which program developers "record some real speech, cut this up into small pieces, and then recombine these to form 'new' speech" (Taylor, 2009, p. 3). The degree to which the synthetic speech is natural-sounding, or its naturalness, is an important criterion when evaluating synthetic speech. However, as Taylor says, "testing in TTS is not a simple or widely agreed-on area" (p. 522).

In addition to the naturalness of synthetic speech, its intelligibility and comprehensibility are important aspects of the speech quality. The intelligibility is determined by "the ability of a listener to decode the message from the speech" (Taylor, 2009, p. 48). Therefore, it is often measured by means of oral or written reproduction tasks after the examinee is made to listen to the speech (e.g., Alamsaputra, Kohnert, Munson, & Reichle, 2006; Axmear et al., 2005; Greene, 1986; McNaughton, Fallon, Tod, Weiner, & Neisworth, 1994; Reynolds, Bond, & Fucci, 1996). On the other hand, comprehensibility, which seems to have received less attention in studies, as discussed by some researchers (e.g., Drager & Reichle, 2001; Reynolds, Isaacs-Duvall, Shward, & Rotter, 2000), is a measure of how much of the message the listener understands. Comprehensibility is different from intelligibility: not all spoken words have to be recognized to comprehend a speaker's intent (Taylor, 2009). Since comprehension comprises many levels, it can be measured in many ways (e.g., true or false questions, a content recall task, or a word/sentence verification task), yet the simplest method may be to administer comprehension questions to listeners (e.g., Sydeserff et al., 1992).

Research Findings and Concerns on Synthetic Speech

Based on a comprehensive review on studies related to the field, Winters and Pisoni (2003) drew seven conclusions about synthetic speech. First, synthetic speech is generally less

intelligible than natural speech, especially in noisier environment with harder tasks and when older synthetic programs are used. Second, synthetic speech requires longer time to process the meaning. Third, listeners are more likely to rely on semantic information when hearing synthetic speech than when hearing natural speech. Fourth, synthetic speech is generally less comprehensible than natural speech even when both speech are equally intelligible, though the gap could be filled if learners can apply off-line strategies to compensate for the lack of on-line comprehension. Fifth, in general, listeners' perception of synthetic speech will get better as they practice listening to it. Sixth, perceptive ability of synthetic speech may vary with the listener's characteristics such as age, physical disabilities, and language background. Seventh, many factors are involved in the evaluation of TTS synthetic speech, yet excellence in sentence-level prosody seems to be most influential in comprehending and perceiving synthetic speech.

Most of these conclusions were obtained through studies on native listeners of English, but it seems plausible that they also hold for non-native listeners. Yet, further research is still necessary to ensure the use of synthetic speech in listening comprehension tests for EFL learners for two reasons. First, compared to the number of studies focusing on synthetic speech intelligibility for non-native listeners (i.e., Alamsaputra et al., 2006; Axmear et al., 2005; Greene, 1986; Reynolds et al., 1996; Venkatagiri, 2005), fewer studies shed light on its comprehensibility. Among the few studies, Jones, Berry, and Stevens (2007) investigated the roles of noise (i.e., multi-talker babble), speech rates (i.e., 155 wpm and 178 wpm), and language background (i.e., native or non-native) for synthetic speech comprehension. The result was that only the main effect of speech rate was significant with no interactions, indicating that there was no difference in comprehensibility between native and non-native listeners. However, the non-significant effects of the other two variables may be due to the fact that the test was composed of true or false questions, which allowed even poorer listeners to answer correctly by guess. Furthermore, this study did not focus on the effect of speech type (i.e., natural or synthetic) on comprehensibility. Regarding speech type comparison, there is a study (O'ki, 2010) that assigned a comprehension test recorded in both natural and synthetic speech and that revealed significant score difference between the two speech types. However, the synthetic speech used in this study was not made by the latest TTS synthesis program. Thus, further research using a more recent program is necessary to reexamine comprehensibility of synthetic speech.

Second, all of the non-native studies listed so far except for that of O'ki (2010) sampled populations quite dissimilar to learners of English in Japan; that is, the participants are either residents or students who have lived in an English speaking country for a certain period of time (i.e., mean residence in an English-speaking country in these studies ranges between about 3 and 10 years). For them, English is important means of daily communication and they have much easier access to a variety of aural English. This kind of input-rich environment is crucial for language learners to improve English perceptive skills, since research has shown that learners need to hear as many varieties of speech as possible to become able to distinguish between phonemic differences that do not exist in their mother tongue (Lively, Logan, & Pisoni, 1993). For this reason, their proficiency levels measured by perceptive skills would be much higher than those of most EFL learners in Japan, and this variable needs to be

considered in research on EFL learners. Yet, little is known about the role of the learner's proficiency on comprehensibility of synthetic speech.

On the basis of these findings and issues, two studies (Study 1 and Study 2) were conducted to find answers to the following three research questions (RQs):

RQ1. Are listening tests created by the TTS synthetic speech program as comprehensible as listening tests in human speech are?

RQ2. Does the TTS synthetic speech sound as natural as human speech does to EFL learners?

RQ3. Does the learner's proficiency level affect the comprehensibility and naturalness of TTS synthetic speech?

Both RQ1 and RQ2 were investigated in Study 1 and Study 2, while RQ3 was examined only in Study 2. With regard to RQ2, although it is assumed that the naturalness of speech can be judged better by native speakers, in this study, the intuitive judgment of non-native speakers was considered to be more appropriate because it concerns relative naturalness between synthetic and human speech among EFL learners, that is, the users of synthetic speech materials.

Study 1

Method

Participants. The participants were 29 undergraduate freshmen who were majoring in economics (7 males and 22 females), and were in a year-round TOEIC preparatory course. To counterbalance the presentation order of natural and synthetic speech, the participants were randomly divided into two groups (Group 1-1, n = 15; and Group1-2, n = 14). Based upon their TOEIC IP listening scores, the two groups possessed equal listening ability skills, $t(26)$ = -0.03, p = .974, d = 0.01.

Materials for the Listening Test. Four passages (Passages A to D) were selected from a prep book for the National Center Test (see Appendix). Each passage was repeated twice by native speakers of English, and, after each replay, there was a 20-second pause to answer three multiple-choice comprehension questions printed on the test sheet. Since Passage B was rather long, several sentences were deleted from it while preserving the consistency of the story. Overall, the four passages were similar in length and difficulty as shown in Table 1.

Table 1

Difficulty of the Four Passages Used in the Listening Test

Passage	Topic	FKGL	FKRE	Letters per Word	Words per Sentence	Total Words	Length (sec)	Rate (wpm)
A	Library	6.3	72.4	4.2	12.9	194	71	163.9
B	Museum	9.9	53.5	4.8	16.5	199	85	140.5
C	Speech	7.9	62.9	4.5	13.9	195	75	156.0
D	UK Family	8.2	56.5	4.5	12.3	198	81	146.7

Note. FKGL = Flesch-Kincaid Grade Level; FKRE = Flesch-Kincaid Reading Ease.

Next, synthetic speech was produced from these passages by using the text-to-speech

(TTS) software, *Globalvoice English ver.2*, to compare it with original human speech. In creating the synthesized speech, variables such as gender of the voice, number of replays, speech rate, and length of pauses were adjusted to match the original speech as faithfully as possible.

Procedures. The listening test consisted of three sections (see Table 2). In Section 1, the participants listened to the test instructions in Japanese and a sample dialogue of about 10 sentences in English, both of which were made by the synthesis program. This section aimed not only at having the participants adjust the volume but at familiarizing them with synthetic sound. The latter purpose was based on the study that listeners' perception of synthetic speech significantly increased after the training of even less than 10 sentences (Rounsefell, Zucker, & Roberts, 1993).

Table 2
Procedures of the Listening Test and the Presentation Order of the Two Speech Types

	n	Section 1	Section 2	Section 3
			Passages A and B	Passages C and D
Group 1-1	15	Instructions	Natural	Synthetic
Group 1-2	14	Instructions	Synthetic	Natural

In Sections 2 and 3, Group 1-1 listened to Passages A and B with natural voice and Passages C and D with synthetic voice, answering three comprehension questions for each passage. In order to minimize the effect of the presentation order of speech types, Group 1-2 started with synthetic speech followed by natural speech. The reason why one section has only two speeches with three questions each was that this study was prompted partly by the researchers' interest in the usability of synthetic speech for the SRST, a speaking test that has been found to be reliable in the case of using two retelling passages of such length and number of questions (Koizumi & Hirai, 2010). The internal-consistency reliability of the four passages using the split-half method (Passages A and B, and Passages C and D) was .70, which was high considering the small number of passages.

At the end of Sections 2 and 3, the participants responded to a short questionnaire that asked how much they thought (1) the rhythm and intonation of the speech was natural ('Prosody'), (2) each word was pronounced correctly ('Segmental'), (3) the speech rate was moderate ('Speech Rate'), and (4) the content was understandable ('Content'), based on a 5-point Likert scale from 1 'strongly disagree' to 5 'strongly agree.' For items (3) and (4), participants were asked to examine if difficulty of the materials was balanced between the two speech types. In addition to these four items, after the test was completed, without being told which speech type students listened to, they were asked to answer which section passages they felt comfortable listening to ('Preference'). Their choices were later categorized into either 'natural,' 'synthetic,' or 'the same.'

Scoring and Data Analysis. The following three kinds of analyses were employed. First, since the listening test had six comprehension questions in each speech type, a maximum of six points were awarded, and the mean scores of the two speech types were compared using a

paired *t* test. Second, participants' perceptions regarding 'Prosody,' 'Segmental,' 'Speech Rate,' and 'Content' aspects, obtained from the questionnaire, were also compared between the speech types using paired *t* tests. Finally, a chi-square test was employed to examine the difference in their speech preference over the three categories (natural, synthetic, or the same).

Results

Comprehensibility of the Two Speech Types. The mean listening scores on the two speech types were nearly the same: the mean score of the natural speech was 3.59 (SD = 1.30) and that of the synthetic speech was 3.62 (SD = 1.29). According to the paired *t* test, the scores of the two speech types were not significantly different, $t(28) = -0.11$, $p = .911$, $d = 0.02$. In addition, if we look at Figure 1 that illustrates mean passage scores with the 95% confidence interval (CI) error bars, the CI ranges largely overlap between the two speech types in each passage. This indicates no significant difference between the natural and synthetic speech comprehension in any of the passages, though some difference in passage difficulty appeared (e.g., Passage C seems to be more difficult than Passage B). Thus, the comprehensibility of synthetic speech made by *Globalvoice English ver.2* was similar to the human speech, which may be an answer to RQ1.

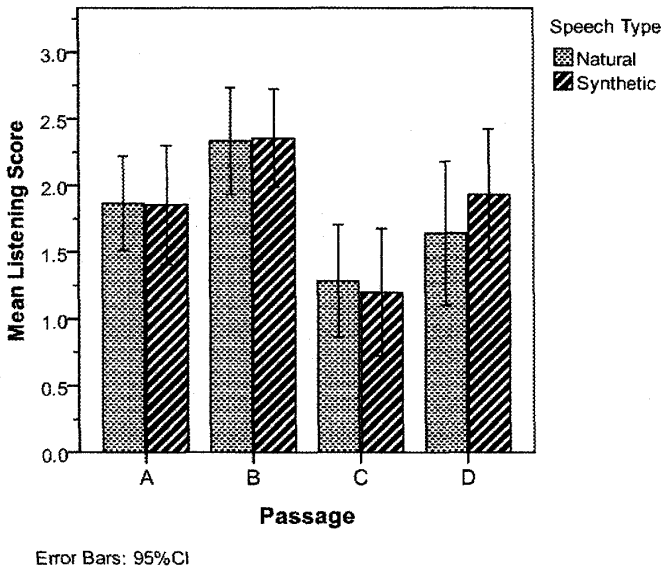

Figure 1. Listening scores of the four passages in two speech types

Naturalnesss of the Two Speech Types. The result where both speech types were equally comprehensible seems to reflect participants' preference of the speech type. Specifically, 16 (55.2%) participants preferred natural speech, whereas only 7 (24.1%) voted for synthetic speech and 6 (20.7%) for the 'same.' It is notable to say that the total of 13 (= 7 + 6) participants, which was nearly half of the whole, did not particularly prefer human speech. The chi square revealed no significant difference in number between those who preferred natural speech and those who did not (i.e., those who chose either 'synthetic' or 'the same'), $\chi^2(1, N = 29) = 0.31$, $p = .577$, demonstrating that natural speech is not always favored by EFL

listeners. However, as Table 3 shows, the mean ratings on Prosody (3.86) and Segmental (3.62) aspects of the natural speech were significantly higher than those of the synthetic speech, showing that the participants considered human speech to be superior in respect of naturalness of rhythm, intonation, and pronunciation of each word. Nevertheless, these perception gaps amount to such a small degree and would not make any difference in comprehension of the two speech types, as far as 'speech rate' and difficulty of 'content' are moderate, which is implied by the moderate ratings on these aspects in both speech types.

Table 3
Participants' Ratings on the Four Items and the Results of T Tests (N = 29)

Item	Natural		Synthetic		*t*	*d*
	M	*SD*	*M*	*SD*		
1. Prosody	3.86	1.06	3.24	1.21	3.00**	0.55
2. Segmental	3.62	1.05	3.28	1.10	2.07*	0.32
3. Speech Rate	3.41	1.05	3.14	1.03	1.44	0.26
4. Content	3.28	0.80	3.03	0.91	1.57	0.29

Note. $*p < .05$., $**p < .01$.

Limitations of Study 1. The overall results may suggest the applicability of synthetic speech to EFL listening materials; however, it might be too hasty to conclude because of the following two limitations. First, most students in Study 1 had participated in another study on synthetic speech that was created by another TTS software two months before this study, thus their prior experience might have familiarized themselves with listening to the synthetic speech. Yet, this experience effect is still doubtful because transferability of experience in one program to another has not been studied before.

Second, the participants of this study were homogeneous with low to intermediate levels of English, so we are still not sure how synthetic speech will work for more proficient learners. Poor listeners, who often lack the ability to detect the segments between words in connected speech (Field, 2008), are likely to have difficulty in understanding naturally connected authentic speech, so that they might not be able to comprehend natural speech better. On the other hand, good listeners, who are supposed to have better perceptive skills, may find it even more comfortable to hear natural speech than unnatural speech, so that they might perform better on natural speech, or with their good perceptive ability that makes up for the acoustic deficiency of synthetic speech, they might perform equally well on both speech types. This is probably why the speech type effect was not found among native listeners in the studies by Alamsaputra et al. (2006) and Axmear et al. (2005). Hence, learners' listening skills, especially their speech perception, should influence their comprehension of the two speech types. To compensate these limitations, the following Study 2 was conducted.

Study 2

In addition to the research questions examined by Study 1, Study 2 probes whether students with different proficiency levels perform differently between the two speech types.

Method

Participants and Procedures. The participants in this study were 75 university students (68 freshmen, 5 sophomores, and 2 seniors; 26 males and 49 females) of various proficiency levels. They were from two universities and were majoring in humanities or English education. Since they had not listened to synthetic speech before, their scores should not have been influenced by prior exposure.

As shown in Table 4, the overall procedure of the study was the same as Study 1 (Sections 1-3), except for the following two points. First, to measure participants' perceptual skills, a dictation test was administered after the listening comprehension test (Section 4). Second, the four passages were reordered to make sure that presentation order would not affect the design of the study. As in Study 1, to counterbalance the presentation order of natural and synthetic speech, students at each university were divided into two groups with an equal English proficiency, based upon an in-house proficiency test at one university, $t(30) = -0.05$, $p = .959$, $d = 0.02$ and on TOEFL ITP examination at the other university, $t(38) = 0.83$, $p = .413$, $d = 0.26$. Then, half of the participants in each group were formed as Group 2-1. The two sub-groups listened to Passages A and D with the human voice and Passages C and B with the synthetic voice. The other sub-groups, formed as Group 2-2, started by listening to Passages A and D with synthetic voice followed by Passages C and B with the natural voice (see Table 4). The internal-consistency reliability of the four passages using the split-half method (Passages A and D, and Passages C and B) was .53, which was moderate considering the small number of passages.

Table 4
Procedures of Study 2

	n	Section 1	Section 2	Section 3	Section 4
			Passages A and D	Passages C and B	
Group 2-1	40	Instructions	Natural	Synthetic	Dictation
Group 2-2	35	Instructions	Synthetic	Natural	Dictation

The dictation test included 12 items recorded in both natural and synthetic speech, each of which required students to write four to seven words. The speech presentation order was counterbalanced between the two groups; that is, Group 2-1 first listened to a half of the items with natural voice and then worked on the other half with synthetic voice, while Group 2-2 did the same in the opposite order. Each dictation sentence was repeated twice.

Scoring and Data Analysis. In the dictation test, there were a total of 70 words and one point was given to each word spelled correctly. The internal consistency was sufficiently high at Cronbach's $\alpha = .84$. Based on the percentage of correct words, the participants were labeled as either Upper ($n = 32$, $M = 70.67\%$, $SD = 6.87\%$) or Lower ($n = 43$, $M = 50.90\%$, $SD = 8.84\%$) proficiency group. Eight participants marked exactly 60.0%, thus this score was determined as the cut-off point for the lower group and the eight participants were all classified into the lower group. A *t* test shows the means of these two groups were significantly different, $t(73) = 10.50$, $p < .001$, $d = 2.50$.

To investigate the influence of the proficiency level on the comprehension of the natural and synthetic speech, a two-way ANOVA in a 2 × 2 (Proficiency [Upper, Lower] × Speech [Natural, Synthetic]) design was conducted. The questionnaire was also analyzed using two-way ANOVAs and chi-square tests.

Results

The Comprehension Test. The descriptive statistics of the mean scores of the two proficiency groups are summarized in Table 5. The participants of both proficiency groups tended to score better on natural speech than on synthetic speech. According to the results of the two-way ANOVA, no significant interaction between Proficiency and Speech was observed, $F(1, 73) = 0.23$, $p = .631$, $\eta_p^2 = .093$, suggesting that both lower and upper proficiency groups performed similarly on both speech types. However, significant main effects of Speech, $F(1, 73) = 7.49$, $p < .01$, $\eta_p^2 = .093$, and Proficiency, $F(1, 73) = 7.89$, $p < .01$, $\eta_p^2 = .097$, were found. These main effects indicate that, contrary to the result of Study 1, natural speech was more comprehensible than synthetic speech, and that the upper proficiency group performed better than the lower proficiency group.

Table 5
Listening Comprehension Scores of the Two Proficiency Groups

	n	Natural		Synthetic	
		M	*SD*	*M*	*SD*
Upper	32	4.44	1.16	3.91	1.09
Lower	43	3.70	1.41	3.33	1.19
Total	75	4.01	1.35	3.57	1.18

Since natural speech was easier to comprehend than synthetic speech, which was different from the result in Study 1, we further investigated passages to find which passages caused the significant disparity between the speech types. As shown in Figure 2, only the error bars in Passage A did not overlap between the speech types, which were found to be significantly different, $t(63) = 2.64$, $p < .05$, $d = 0.61$. This may be because, as explained with Table 4, Passage A was the first synthetic speech that Group 2-2 listened to, and they probably had not become familiar with synthetic voice yet. A similar phenomenon was seen for Passage C, of which Group 2-1 worked on the synthetic speech for the first time and the mean difference almost reached the significant level ($p = .077$).

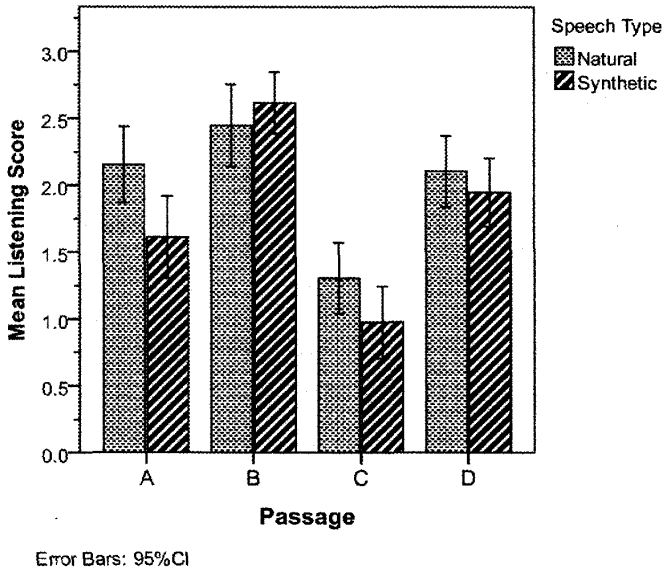

Figure 2. Mean scores of the two proficiency groups in each speech type

The inconsistency of the results between Study 1 and Study 2 could be due to the difference in participants' experience in listening to synthetic speech. As anticipated in the limitation of Study 1, most participants in Study 1 had listened to synthetic speech before, which may have affected the comprehension of the synthetic speech of this study.

Participants' Responses to the Questionnaire. As to the results on the questionnaire, two groups' ratings on the four items are summarized in Table 6. Overall, the upper group gave higher rating to each item, and the mean difference between the speech types is consistently small in all the items of both proficiency groups. To test if there were differences between the two proficiency groups and between the two speech types, a two-way ANOVA in a 2 × 2 (Proficiency [Upper, Lower] × Speech [Natural, Synthetic]) was performed for each item.

Table 6

Two Proficiency Groups and Their Mean Ratings on the Four Items

	Upper				Lower			
	Natural		Synthetic		Natural		Synthetic	
Item	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1. Prosody	4.25	0.84	4.19	0.82	3.86	1.13	3.53	1.22
2. Segmental	4.19	0.90	4.22	0.83	3.84	1.04	3.58	1.14
3. Speech Rate	3.63	1.10	3.41	1.07	3.19	1.22	3.23	1.27
4. Content	3.90	0.94	3.55	0.83	3.30	1.08	3.35	0.97

Table 7

Two-way (Proficiency and Speech) Analyses of Variance for the Four Items

Variable and source	*MS*	*F*	*p*	η_p^2
1. Prosody				
Proficiency	9.96	6.49	< .05	.082
Speech	1.38	2.16	.146	.029
Proficiency × Speech	0.64	0.99	.322	.013
2. Segmental				
Proficiency	8.95	6.22	< .05	.078
Speech	0.46	0.81	.370	.011
Proficiency × Speech	0.76	1.33	.253	.018
3. Speech Rate				
Proficiency	3.44	1.64	.205	.022
Speech	0.27	0.40	.529	.005
Proficiency × Speech	0.65	0.95	.333	.013
4. Content				
Proficiency	5.50	4.51	< .05	.060
Speech	0.77	1.14	.289	.016
Proficiency × Speech	1.33	1.97	.165	.027

df = 1, 73.

As summarized in Table 7, significant interaction was not found between Proficiency and Speech in any of the four aspects, nor was a significant main effect of Speech. Significant effect of Proficiency in three aspects indicates that the upper proficiency group was more certain of their ratings than the lower group. These results indicate that in all the aspects, irrespective of the proficiency level, learners' perceptions were not different between the two speech types. This implies that the synthetic speech was perceived as natural as human speech. In particular, very high ratings on Prosody (4.19) and Segmental (4.22) aspects of the synthetic speech by the upper group may indicate that even high proficiency learners perceived the synthetic speech to be natural with strong confidence.

As for participants' preference over the speech types (see Table 8), on the whole, the majority of the learners (n = 39 [60.9%]) preferred natural speech to synthetic speech, yet the rest of the students (n = 12 and 13 [39.1%]) did not particularly prefer natural speech, and the difference in number did not reach the significant level (p = .080).

However, this tendency turned out to be slightly different between the upper and lower groups. A significantly larger percentage of the upper proficiency learners (n = 17 [70.8%]) preferred the natural speech, and the rest of them (n =7 [29.2%]) felt that synthetic speech was more natural than or as natural as the human speech, and the difference was statistically significant, $\chi^2(2, N = 24) = 4.17$, $p < .05$. On the other hand, 22 learners (55.0%) in the lower group voted for natural speech, but as many as 18 learners (45.0%) either preferred synthetic speech (n = 10 [25.0%]) or showed no preference to natural speech (n = 8 [20.0%]), and this difference was not significant, $\chi^2(1, N = 40) = 0.40$, $p = .527$. Thus, similar to the result of Study 1, nearly half of the lower proficiency learners accepted synthetic speech.

This high preference of synthetic speech over human speech among the lower proficiency learners was observed in the rating on Speech Rate and Content aspects. Even though there

was no significant difference in these aspects between the two speech types, it is interesting to note that the low proficiency group gave slightly higher ratings to synthetic speech than natural speech on these aspects. Thus, more students in the lower proficiency group may feel it easier to listen to synthetic speech rather than natural speech.

Table 8
A Crosstab on the Two Groups' Responses to the Preference Test

Proficiency	Preference			Total
	"Natural"	"Synthetic"	"Same"	
Upper	17 [70.9%]	2 [8.3%]	5 [20.8%]	24 [100.0%]
Lower	22 [55.0%]	10 [25.0%]	8 [20.0%]	40 [100.0%]
Total	39 [60.9%]	12 [18.8%]	13 [20.3%]	64 [100.0%]

Note. There were 11 missing cases; Numbers in the brackets show percentages in the same proficiency group.

General Discussion and Conclusion

This study started due to the researchers' interest in whether synthetic speech which was created by the latest TTS synthetic program could replace natural human speech for its use in an EFL comprehension test. It would be very helpful if English teachers and researchers could easily convert written texts into speech that is as natural and comprehensible as human speech. To this end, we had three RQs and conducted two studies.

Comprehensibility of Synthetic Speech and the Role of Training

Concerning RQ1 (whether synthetic speech is as comprehensible as natural speech), Study 1 resulted in no significant difference in listening test score, but Study 2 revealed the significant difference. The degree of familiarity to synthetic speech might have caused the different results. Most students in Study 1 had experienced listening to synthetic speech attentively once before, while students in Study 2 had not. In fact, in Study 2, the scores of the synthetic speech they first listened to were lower than those of the natural speech, and this tendency was not observed in Study 1. To support this point, there is a study reporting that the effect of synthesized speech training may last for more than six months (Winters & Pisoni, 2003). Thus, sufficient practice in listening to synthetic speech before conducting a test seems to be important.

Besides the effect of prior exposure to synthetic voice, another factor that influences the comprehension of synthetic speech might be speech rate as pointed out by Jones et al. (2007) and content of speech. Although optimal speech rate of low proficiency listeners is generally lower than that of high proficiency listeners (e.g., Hirai, 2010), the lower proficiency group as well as the upper felt the speech rate and the difficulty of content of both types of speech to be moderate. Thus, the speech rate and the difficulty of the passage used in this study were equally and appropriately adjusted to both speech types, and thus did not seem to affect the

comprehension of the synthetic speech nor natural speech.

Naturalness of Synthetic Speech

Regarding RQ2 (whether synthetic speech sounds as natural as human speech), the participants' responses to the four items in the questionnaire were also different in Study 1 and Study 2. In Study 1, human speech was significantly perceived better than synthetic speech in such aspects of Prosody and Segmental, whereas students in Study 2 gave similar ratings to both types of speech. However, the ratings on the phonological aspects (i.e., Prosody and Segmental) of synthetic speech were relatively high in both studies. In particular, the upper proficiency learners in Study 2 rated an average of 4.00 on the five-point scale on Prosody and Segmental aspects. This means that synthetic speech was perceived almost as natural as the human speech. In addition, many learners, especially the lower proficiency learners, in both studies did not particularly prefer natural speech. This may imply that synthesized speech can be used for teaching and testing EFL learners, and this does not necessarily lead to distraction of learners' attention or loss of motivation due to the quality of synthetic speech. Thus, though the applicability of synthetic speech may depend on which synthetic software we use, it seems that today's TTS synthesis technology has reached the level where we could produce speech quite close to natural human speech.

The Role of Learners' Proficiency on Comprehensibility and Naturalness of Synthetic Speech

In regard to RQ3 (whether proficiency is related to the comprehensibility and naturalness of the speech type), mixed results were observed in the three tasks (i.e., the comprehension test, the questionnaire on the four aspects, and the preference question). As for the comprehension test, there was no significant interaction between proficiency and speech type. In other words, the upper level group consistently comprehended both natural and synthetic speech better than the lower group. This indicates that comprehension test made by synthetic speech can also be used to distinguish learners' proficiency levels. Similarly, no interaction between the variables was found in any of the four aspects of the questionnaire, indicating that both groups evaluated naturalness of the two speech types equally.

However, when it comes to preference of speech type, proficiency factor was evident. When students were not told which speech was synthetic or a human voice, approximately 70% of the upper proficiency group considered human speech to be more comfortable to listen to, whereas 55% of the lower proficiency group chose human speech and the remaining 45% of them preferred synthetic speech or did not mind about either speech type. The latter tendency also appeared in Study 1, in which the proficiency level of the participants was similar to that of the lower group in Study 2.

There could be two reasons for this preference difference between the two proficiency groups. First, lower proficiency listeners may feel it easier to listen to synthetic speech because it is different from natural or authentic speech where phonetic variation is rich and fluent, whereas synthetic speech is read at a constant speed in all sections of the speech, and each word is regularly segmented. In this sense, synthetic speech might be less natural but could be more intelligible, as Taylor (2009) points out that "there is an inverse correlation

between naturalness and intelligibility" (p. 48). Second, the lower level group could not differentiate the phonetic quality of natural speech from that of synthetic speech, since they did not have enough mental capacity to listen to the sound while focusing on the meaning of the speech. This claim can be supported not only by their low comprehension scores in the two studies, but by their wider *SD*s in Prosody and Segmental aspects especially in Study 2, which could be a sign of their insecure judgment.

Pedagogical Implications and Limitations

Overall, positive results showed that synthetic speech may be used for both upper and lower proficiency groups in conducting listening tests, since it enables the teacher to discriminate between lower and upper proficiency groups as was used in the human speech. In addition, since recent TTS synthesis technology produces quality speech that is similar to human speech, various effective uses of synthetic speech can be suggested. For example, lower proficiency students who feel that it is too difficult to comprehend real authentic speech may first practice listening to synthetic speech which is converted from a script and then work on the original authentic material later. The program will also enable instructors and researchers to produce various materials for SRST or listening-related tasks such as shadowing practice. The teacher can also transform students' speech scripts into synthetic speech in order for them to practice pronunciation individually before, for example, a speech contest. In all of these cases, a great part is that Japanese instructors and researchers do not have to take the extra step to ask native speakers to record texts every time listening materials are needed. Thus, the use of today's TTS synthetic speech software has a great potential for non-native teachers and researchers of English.

As to limitations and future research, in the studies, only two passages were used for each speech type, and only three comprehension questions were assigned to each passage. This was because an investigation was made to see whether synthetic speech materials of intended length, difficulty, and number of comprehension questions could be used for the SRST. In this manner, the results may not be applicable to wider contexts of English education. For example, we are not sure whether more difficult or more colloquial materials can be replaced by synthetic ones. In this respect, we need to conduct a comparative study under such different conditions.

A second research area is the effect of practice by students on the comprehensibility of synthetic speech. If the score gaps in the passages that appeared in Study 2 were due to participants' lack of experience in listening to synthetic speech, instructions using synthetic voice might not be sufficient. Thus, further study is necessary to examine to what extent learners should practice before taking the test to get accustomed to synthetic speech.

Acknowledgements

We are grateful to the anonymous reviewers for their insightful comments on an earlier version of this paper and partial support by Grant-in-Aid for Scientific Research (KAKENHI) (C) (23520744).

References

Alamsaputra, D. M., Kohnert, K. J., Munson, B., & Reichle, J. (2006). Synthesized speech intelligibility among native speakers and non-native speakers of English. *Augmentative and Alternative Communication, 22,* 258–268.

Axmear, E., Reichle, J., Alamsaputra, M., Kohnert, K., Drager, K., & Sellnow, K. (2005). Synthesized speech intelligibility in sentences: A comparison of monolingual English-speaking and bilingual children. *Language, Speech, and Hearing Services in Schools, 36,* 244–250.

Azuma, J. (2010). Impact of TTS technology on foreign language teaching: New horizons of multimedia teaching material development. *Ryutsu Kagaku Daigaku Kyoiku Koudoka Suishin Center Kiyou, 6,* 1–11.

Drager, K. D. R., & Reichle, J. E. (2001). Effects of age and divided attention on listeners' comprehension of synthesized speech. *AAC (Augmentative and Alternative Communication), 17,* 109–119.

Field, J. (2008). *Listening in the language classroom.* Cambridge: Cambridge University Press.

Greene, B. G. (1986). Perception of synthetic speech by nonnative speakers of English. In *Proceedings of the Human Factors Society-30th Annul Meeting* (pp. 1340–1343) Bloomington, IN: Speech Research Laboratory.

Globalvoice English (Version 2) [Computer software]. Tokyo: Hoya.

Hirai, A. (2010). *L2 listening and reading fluency: A comparative study of English with diverse L1 backgrounds.* Berlin: Lambert Academic Publishing.

Hirai, A., & Koizumi, R. (2009). Development of a practical speaking test with a positive impact on learning using a story retelling technique. *Language Assessment Quarterly, 6,* 151–167.

Jones, C., Berry, L., & Stevens, C. (2007). Synthesized speech intelligibility and persuasion: Speech rate and non-native listeners. *Computer Speech and Language, 21,* 641–651.

Koizumi, R., & Hirai, A. (2010). Exploring the quality of the Story Retelling Speaking Test: Roles of story length, comprehension questions, keywords, and opinions. *ARELE (Annual Review of English Language Education in Japan), 21,* 211–220.

Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *Journal of the Acoustical Society of America, 94,* 1242–1255.

McNaughton, D., Fallon, K., Tod, J., Weiner, F., & Neisworth, J. (1994). Effect of repeated listening experiences on the intelligibility of synthesized speech. *AAC (Augmentative and Alternative Communication), 10,* 161–168.

Oki, T. (2010a). Investigating the role of shadowing for facilitating bottom-up processing. *Tsukuba Review of English Language Teaching (Tsukuba Eigo Kyoiku), 31,* 1–22.

Oki, T. (2010b). The role of latency for word recognition in shadowing. *ARELE (Annual Review of English Language Education in Japan), 21,* 51–60.

O'ki, T. (2010). Applicability of synthesized speech for EFL listening tests. *The Hakuoh University Journal (Hakuoh Daigaku Ronshu), 25(1),* 195–209.

Reynolds, M. E., Bond, Z. S., & Fucci, D. (1996). Synthetic speech intelligibility: Comparison of native and non-native speakers of English. *AAC (Augmentative and Alternative Communication), 12,* 32–36.

Reynolds, M. E., Isaacs-Duvall, C., Shward, B., & Rotter, M. (2000). Examination of the effects of listening practice on synthesized speech comprehension. *AAC (Augmentative and Alternative Communication), 16,* 250–259.

Rounsefell, S., Zucker, S. H., & Roberts, T. G. (1993). Effects of listener training on intelligibility of augmentative and alternative speech in the secondary classroom. *Education and Training in Mental Retardation, 28,* 296–308.

Sydeserff, H. A., Caley, R. J., Israd, S. D. Jack, M. A., Monaghan, A. I. C., & Verhoeven, J. (1992). Evaluation of speech synthesis techniques in a comprehension task. *Speech Communication, 11,* 189–194.

Taylor, P. (2009). *Text-to-speech synthesis.* New York, NY: Cambridge University Press.

Venkatagiri, H. S. (2005). Phoneme intelligibility of four text-to-speech products to nonnativespeakers of English in noise. *International Journal of Speech Technology, 8,* 313–321.

Winters, S. J., & Pisoni, D. B. (2003). Perception and comprehension of synthetic speech. In *Research on spoken language processing progress report No. 26* (pp. 95–138). Bloomington, IN: Indiana University, Speech Research Laboratory.

Daigaku Nyushi Center Shiken: Jissen Mondaishu English. (2010). Tokyo: Yoyogi Library.

Appendix. *An Example of the Listening Materials.*

Passage A : *Good afternoon, ladies and gentlemen, and welcome to the tour of the Kyoto National Library on this lovely afternoon. My name is Taka Jingu and I'm your guide for today. Could I please see your tickets for the guided tour? I'd also like to remind you that any tickets bought today do not include a visit to the reading rooms. I'm afraid we don't do visits on Fridays, or any weekday, so as not to disturb the readers. But if you do want to see those rooms, the only day there are tours is on Sundays. So I don't want anyone to be disappointed about that day. OK? Thank you. Right. We'll start with a short introduction. As many of you know, this is Japan's National Library and you can see that this is a magnificent modern building. Construction started in 1998 and it was completed in the last year of the twentieth century. This library holds more than fifty million books. But this isn't a public library. You can't just come in and join and borrow any of the books. Access to the collection is limited to those involved in carrying out research.*

Q1. Why can't today's tourists visit the reading rooms?

1. Only those who bought their tickets beforehand can visit them.
2. The library is too big to tour in a day.
3. The reading rooms are not open to tourists on weekdays.
4. There are too many tourists.

Q2. What does the library look like?

1. It's a high-tech building although it looks old.
2. It's a newly constructed modern building.
3. It's an old-fashioned building with modern equipment.
4. It's a traditional Japanese-style building.

Q3. What is the purpose of this library?

1. To help those who research something.
2. To hold as many books as possible.
3. To lend books to as many people as possible.
4. To preserve Japan's traditional culture.