JACET Journal 61 (2017) 149–164

Evidence of Effects of Text-to-Speech Synthetic Speech to Improve Second Language Learning

MATSUDA, Noriko
Osaka Institute of Technology

Abstract

In recent years, the number of applications of synthesized speech software in the English language classroom has increased. The use of text-to-speech (TTS) synthesized technology is expected to shed light on the dearth of spoken English input for learners in Japan. However, it still remains unclear whether synthetic speech has similar learning effects to natural human speech for effectively learning a second language. This study investigated auditory word priming in 80 Japanese English as a foreign language (EFL) learners. Vocal repetition data was used to measure perceptual learning. The results revealed that by focusing more on a perceptual dimension, (1) there was significantly greater learning when natural human speech was used. However, when synthetic speech was used, (2) the lower proficiency group showed a sufficient level of learning. Also, when synthetic speech was used, (3) significantly greater learning occurred within the higher proficiency group when the focus was more on a higher-level dimension (i.e., a semantic dimension). Further research is required to apply the results to English language education in Japan, although the results support the application of synthetic speech as an effective tool to increase perceptual learning in Japanese EFL students.

***Keywords*:** text-to-speech synthesis, priming experiment, perceptual learning

Introduction

Background

TTS (text-to-speech) synthesizing software that allows teachers and students to freely create foreign speech has an enormous potential to solve the problem of limited second language (L2) input. Several cases of speech synthesis in English-language classrooms have been reported following the rapid advancements in speech synthesis technology in recent years (Azuma, 2010; Kataoka & Ito, 2013). These cases reveal a variety of advantages from the possibility of developing different kinds of speech learning material to broadening educational activities. Adding and editing data to synthetic digital speech is simplified using speech synthesizing software, which may possibly lighten the burden of teachers by eliminating their tasks of contacting native speakers individually, recording, and then editing their voices. In addition, the use of synthesized TTS is not limited to learning activities, as it can also possibly help in research, such as in conducting psycholinguistic experiments, again, because it is easy to control the necessary stimulation. However, despite these various possibilities, there are few studies on the possibilities of using TTS in foreign-language classrooms or comparative

studies to natural human speech (Azuma, 2010; Kashiwagi, Kang, & Ohtsuki, 2008).

To be able to automatically process phonetic input (a conscious process that increases in speed after repeated drills and transitions into an unconscious process), which is the basis of the spoken language process, it is essential for foreign language learners to be able to correctly decipher what they hear in the target language. As a result, there have been a large number of studies in the past ten years in the field of foreign language education in Japan, focusing on the effects of training that facilitates the perceptual process, such as shadowing (a training method wherein learners immediately repeat what they hear) (e.g., Kadota, 2007, 2015; Tamai, 2005). Therefore, understanding the benefits of synthesized TTS for this type of training offers the potential of using synthesized TTS to improve the listening skills of Japanese learners of English. With this background, the researcher conducted a priming experiment to compare and investigate the perceptual learning effects[1] of using synthesized TTS and natural human speech.

Previous Studies on Priming

Priming refers to the influence that processing previous stimuli has on processing subsequent stimuli. A number of experiments on auditory priming where vocabulary has been aurally presented with recorded natural voices have shown that vocabulary heard once could be correctly repeated faster than new vocabulary. Previous studies indicate that the reason for this is that learners memorize the acoustic properties of a voice at the perception level and use the information unconsciously (Trofimovich & Gatbonton, 2006). This represents a learning effect at the perception level. This kind of learning occurs because humans use previous information to carry out their daily routines smoothly and efficiently. Auditory priming is known to be a universal mechanism that aids in language acquisition and, additionally, it has been suggested that this mechanism may work in the acquisition of languages other than the mother tongue (McDonough & Trofimovich, 2009, p. 28).

In studies focused on one's native language, there was no visible difference in the priming effect when listening to vocabulary, whether the focus was on the sound or the meaning of the material presented (McDonough & Trofimovich, 2009; Trofimovich, 2005). However, the few studies that have focused on L2 indicate that contrary to learning one's mother tongue, there is a negative impact on the priming effect based on the person's proficiency when focusing on meaning (Trofimovich, 2005, 2008; Trofimovich & Gatbonton, 2006). These studies explain that "L2 learners may not benefit from repeated experiences with spoken words, at least early in their L2 development or after a relatively brief experience with the L2, when they engage in a meaningful, semantic processing of words" (= no perceptual learning effect) (McDonough & Trofimovich, 2009, p. 30). The subjects of these studies were L2 learners in auditory-input-rich ESL environments. These studies also had various definitions for proficiency. Trofimovich (2008) defined the barometer of proficiency as the length of residence in the country where L2 is the national language, while Trofimovich and Gatbonton (2006) defined it as the degree of pronunciation ability. The auditory priming effect itself can also be seen in studies where subjects were Japanese students in EFL environments dissimilar from other ESL environments (Matsuda, 2013, Sugiura & Hori, 2012). However, the author could not locate detailed studies of the auditory priming effect on EFL learners that considered both

proficiency and focus when students were listening to vocabulary.

Previous Studies on Synthetic Speech

The most popular kind of speech synthesis technology in use today is rule-based speech synthesis. It is known as corpus-based speech synthesis technology based on a large-scale database from natural voices, such as from professional announcers. It "generates synthesized speech by editing the voice waveform segment data and varying it for intonations and such according to synthesis rules established beforehand" (Watanabe, Iwaki, Kaneyasu, & Miki, 2006). This is characterized by speech that feels authentic because it connects fragments of natural human speech. The TTS synthesis software used in this experiment also uses this method.

There is continuing research into intelligibility and comprehensibility in synthetic speech, with the former being of relevance to this study, especially because the study objective is to understand the perceptual learning effect. Multiple studies have been conducted to find contributing factors, such as how age differences in students affect intelligibility (e.g., Drager, Reichle, & Pinkoski, 2010; Pinkoski-Ball, Reichle, & Munson, 2012) or repetition effects (e.g., Koul & Clapsaddle, 2006; McNaughton, Fallon, Tod, Weiner, & Neisworth, 1994; Reynolds & Jefferson, 1999) in one's native language. In addition, speaking or speech rate, noise, and linguistic context have all been presented as factors that influence speech intelligibility (Axmear et al., 2005, p. 245). However, speech rate has been found to be an especially important factor that influences not only intelligibility, but also comprehensibility (Jones, Berry, & Stevens, 2007).

Few studies exist that focus on speech intelligibility in L2. Axmear et al. (2005) assigned repetition tasks to monolingual and bilingual children that revealed that intelligibility was higher for natural voices than synthetic ones and intelligibility of synthetic speech was lower in bilingual children than in monolingual children. Similar results were obtained with adults in a study of Venkatagiri (2005), even though they assigned written and not repetition tasks.

Hirai and O'ki (2011) focused on the comprehensibility of synthetic speech with Japanese learners of English. This study indicated that although comprehensibility among learners tended to be higher with natural human speech, synthetic speech was perceived to be almost the same as natural. Moreover, the "experience effect" influenced the comprehensibility of synthetic speech after hearing the speech once. Despite this, a higher percentage of students with low proficiency (25.0%) preferred synthetic speech compared to students with higher proficiency levels (8.3%). They believe this is due to the fact that "synthetic speech is read at a constant speed in all sections of the speech, and each word is regularly segmented," making it easier for the "lower proficiency listeners" to listen to it (p. 13). They argue that their study shows that synthetic speech can be used for English education.

Based on previous studies of L2 speech intelligibility, the perceptual learning effect can be expected to be greater when using natural human speech rather than synthetic speech especially for students with higher proficiency levels. Also, it is likely that unnatural features of synthetic speech, such as steady reading speed and regular segmentation, will influence the preferences of higher proficiency level students and reduce the perceptual learning effects of synthetic speech.

Studies that investigated the auditory priming effect in Japanese learners of English have shown a learning effect when using both recorded natural human speech (Sugiura & Hori, 2012) and synthetic speech (Matsuda, 2013). Although these studies did not compare both sides, the researcher believes it is possible to compare the learning effect of using both synthetic speech and natural human speech at the perception level by controlling various factors including speech rate.

Research Questions and Hypotheses of the Study

This study investigated auditory word priming to answer the following research questions:

RQ1. Is there a difference in the perceptual learning effects when using natural human speech rather than synthetic speech?

RQ2. Does the perceptual learning effect change based on the learner's focus when listening to speech?

RQ3. Does the perceptual learning effect change with the proficiency level of the learner?

The following hypotheses were constructed based on previous studies on priming and synthetic speech to answer the above research questions:

H1. The perceptual learning effect will be greater with natural human speech than with synthetic speech.

H2. The perceptual learning effect will decrease when using natural human speech and focusing on meaning.

H3. The perceptual learning effect will be greater using natural human speech, but decrease using synthetic speech for learners with high proficiency levels.

Method

Participants

Participants in this study were 80 Japanese learners of English (undergraduate and graduate students). They were divided into two equal groups. One group of participants participated in the experiment that used synthetic speech and the other participated in the experiment that used recorded natural human speech. The participants in these two experiments did not overlap. Table 1 shows data on the participants.

Based on the correct answer rate of the Oxford Quick Placement Test (Oxford University Press et al., 2001) (60 points maximum) and the Oxford Placement Test (Listening Test) (Allan, 2004) (100 points maximum), there was no significant difference in the proficiency of these two groups, F (1, 78) = 1.84, p = .17, η_p^2 = .02.

Materials

Vocabulary groups were created based on controlled familiarity (Yokokawa et al. 2006, 2009), F (3, 68) = 0.19, p = .90, η_p^2 = .008 (Spoken) and F (3, 68) = 0.60, p = .61, η_p^2 = .03 (Written), frequency (British National Corpus or BNC), F (3, 68) = 0.86, p = .46, η_p^2 = .04, number of syllables, F (3, 68) = 0.06, p = .98, η_p^2 = .003, duration, F (3, 68) = 0.03, p = .99, η_p^2 = .001, and initial consonants (the initial consonants of three pairs were uncontrolled) using synthesized TTS speech or natural human speech from native English speakers (see Appendices A and B).

Table 1
English Learning Background of Participants

	Natural (n = 40)		Synthetic (n = 40)	
	M	*SD*	*M*	*SD*
The Oxford Quick Placement Test	30.95	7.14	30.05	5.41
The Oxford Placement Test (Listening Test)	74.67	8.35	71.13	5.60
Age starting English study	11.41	1.72	10.85	2.40
Years of formal English education	11.92	1.33	11.98	1.05
Years of residence in English-speaking countries	0.20	0.96	0.18	0.78
Age	20.90	2.67	19.08	0.80
Self-ratings[a]	4.51	1.22	4.94	1.09
Listening	4.56	1.69	4.65	1.51
Speaking	3.92	1.38	4.35	1.51
Reading	4.97	1.39	5.38	1.39
Writing	4.56	1.48	4.78	1.56

Note. *SD* = standard deviation. [a]Ratings scored on a 10-point scale with 1 = minimum proficiency and 10 = near-native proficiency.

The Globalvoice English Professional version 2.0.1 (HOYA) was used as the speech synthesis software with Kate as the female voice and Paul as the male voice. For the natural human speech, the researcher requested assistance from two native English speakers (one female and one male, citizens of the United States who were both working as English teachers in Japanese universities). Their voices were monaurally digitally recorded in a sound booth using an IC recorder (SONY ICD-SX67) with a microphone (SONY ECM-DS70P) at a sampling rate of 48Hz and quantization at 16 bits. The following method was used to set the speed of speech, a known major factor from previous studies that influences intelligibility and comprehensibility. The English teachers were provided sounds for the vocabulary group created using the above-mentioned speech synthesis software and were asked to repeat each word twice while being conscious of speed. Steps were taken to ensure that there were no significant differences in the speech rate between the recorded speech and the speech from the synthesizing software. All words were checked by the researcher and if the speech rates were different, Praat (Boersma & Weenink, 2012) was used to match them. The researcher also ensured similar volumes for the speech using Praat (mean 73.0 dB).

The experiment was divided into a study phase and a test phase and used rhyme judgment tasks and synonym judgment tasks during the study phase. Participants were expected to focus on the sound of the words during the rhyme judgment task. Specifically, this task used 18 pairs of words, and consisted of determining whether two words rhymed (e.g., remember – November) and pressing the correct key. Participants were expected to focus on the meaning of the words during the synonym judgment task. This task used a new set of 18 pairs of words, and consisted of determining whether two words had similar meanings (e.g., exercise – training) and pressing the correct key. Repetition tasks were used during the test phase. Eighteen words from each pair of the study phase (e.g., 'remember' and 'exercise' from the examples were used and 'November' and 'training' were discarded) and 18 new words,

which did not appear in the study phase, were used during the test phase (see Appendices A and B).

Procedure and Data Analysis

The experiment was conducted on an individual basis using a computer. The words were provided randomly using SuperLab 4.5 (Cedrus Corporation) and participants were asked to respond as quickly and as accurately as possible for all tasks. Practice time was determined before conducting each task in order to ensure that the tasks could be performed smoothly. An example of the task order for a participant is as follows: rhyme judgment task (study phase) → mathematic task[2] → repetition task (test phase) → rest → synonym judgment task (study phase) → mathematic task → repetition task (test phase). The order of the tasks was counterbalanced across the participants to avoid influencing the results.

The analysis target was the reaction time (RT) from the onset of the model sound to the onset of the repetition during the test phase. In accordance with previous studies, the priming effect was determined by whether there was a significant difference in the RTs for previously encountered vocabulary during the study phase and new vocabulary presented during the test phase. The RTs were calculated using Praat and an analysis of variance (ANOVA) was conducted to investigate the priming effects.

The researcher considered whether the focus was on the sound or the meaning during the study phase and investigated how using synthetic speech was different from using recorded natural human speech. In addition, a proficiency-based analysis was added.

Following the example set by previous studies (Trofimovich, 2008), two evaluators, including the author of this paper, selected repetitive data that were deemed errors (the concordance rate of error evaluations was 96.88%). The incorrect data, 4.51% of the data, was excluded from the analysis. The RT data that was two standard deviations (*SD*) away from each participant's mean was substituted by the sum of the mean and 2*SD*s.

Results

Learning Effect When Using Natural Human Speech and Synthetic Speech

Table 2 shows the mean RTs of the repetition task when using natural human speech and synthetic speech. A mixed design 2 (Speech) × 2 (Task) × 2 (Repetition) ANOVA was conducted with Speech (recorded natural human speech or synthetic speech) as a between-subject factor, while Task (rhyme judgment task or synonym judgment task in the study phase) and Repetition (whether the vocabulary was heard once before during the study phase) as within-subject factors. The results of the three-way ANOVA showed a significant three-way interaction, F (1, 78) = 18.58, p < .001, η_p^2 = .19. The simple interaction effect was investigated using a mixed design 2 (Speech) × 2 (Repetition) ANOVA for each Task. There was a significant interaction between Speech and Repetition when the rhyme judgment task was conducted during the study phase, F (1, 78) = 11.24, p = .001, η_p^2 = .13. The results of the simple main effect test revealed a repetition effect for both natural human speech, F (1, 78) = 76.56, p < .001, η_p^2 = .50, and synthetic speech, F (1, 78) = 16.07, p < .001, η_p^2 = .17. There was also a significant interaction between Speech and Repetition when the synonym judgment task was conducted during the study phase, F (1, 78) = 9.45, p = .003, η_p^2 = .11. The results of the

simple main effect test revealed a repetition effect for synthetic speech, F (1, 78) = 41.05, $p < .001$, $\eta_p^2 = .34$, but not for natural human speech, F (1, 78) = 4.24, $p = .09$, $\eta_p^2 = .05$.

The results show a positive priming effect both when a real human voice was used and when synthetic speech was used. However, the statistical results and Figure 1 show that the size of the priming effect changed for the combination of Speech and Task. The perceptual learning effect was greater with recorded human speech when participants focused on the sound of the vocabulary while it was greater with synthetic speech when participants focused on the meaning.

Table 2
Mean RTs (ms) of the Repetition Task

	Natural				Synthetic			
	Rhyme		Synonym		Rhyme		Synonym	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Unrepeated	836.44	106.70	840.26	104.50	824.40	148.81	835.65	145.27
Repeated	806.70	112.70	832.53	109.53	810.60	142.87	811.75	149.73

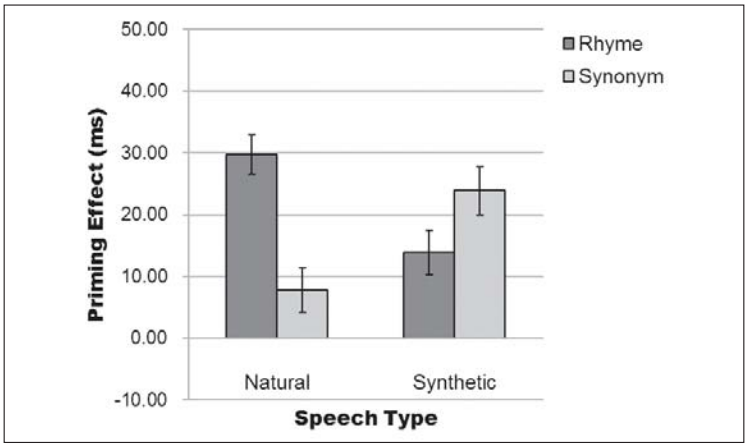

Figure 1. Mean priming effects (ms). The vertical lines indicate standard error.

Proficiency-Based Analysis

To measure the effects of proficiency, the researcher classified these learners into three groups based on the results of the proficiency tests, discarded the middle group and compared the upper- and lower-proficiency groups. The data of the upper- and lower-proficiency groups are shown in Table 3 (natural human speech) and Table 4 (synthetic speech). Significant differences are revealed in the scores for both the Oxford Quick Placement Test and the Oxford Placement Test (Listening Test) at each proficiency level.

Table 3

Two Proficiency Groups and Their English Learning Background When Using Natural Human Speech

	Lower (n = 12)		Upper (n = 12)		F	p	η_p^2
	M	*SD*	*M*	*SD*			
The Oxford Quick Placement Test	27.17	3.49	33.83	6.55	9.69	.005	.306
The Oxford Placement Test (Listening Test)	68.91	6.01	73.42	3.63	4.84	.039	.187
Age starting English study	9.92	3.34	11.75	1.14	3.23	.086	.128
Years of formal English education	11.92	1.62	12.08	0.51	0.12	.738	.005
Years of residence in English-speaking countries	0.59	1.37	0.01	0.03	2.14	.158	.089
Age	18.83	0.58	19.33	0.89	2.68	.116	.108
Self-ratings[a]	4.73	1.40	5.25	1.02	1.09	.308	.047
Listening	4.83	1.85	4.50	1.45	0.24	.628	.011
Speaking	4.58	1.62	4.50	1.62	0.02	.901	.001
Reading	4.83	1.59	6.50	1.00	9.48	.006	.301
Writing	4.67	1.78	5.50	0.90	2.10	.162	.087

Note. *SD* = standard deviation. [a]Ratings scored on a 10-point scale with 1 = minimum proficiency and 10 = near-native proficiency.

Table 4

Two Proficiency Groups and Their English Learning Background When Using Synthetic Speech

	Lower (n = 12)		Upper (n = 12)		F	p	η_p^2
	M	*SD*	*M*	*SD*			
The Oxford Quick Placement Test	24.58	2.68	38.08	4.85	71.21	< .001	.764
The Oxford Placement Test (Listening Test)	68.00	7.41	76.33	7.36	7.63	.011	.258
Age starting English study	11.42	1.24	11.58	2.11	0.06	.816	.003
Years of formal English education	12.00	0.85	12.33	0.65	1.16	.294	.050
Years of residence in English-speaking countries	0.00	0.01	0.14	0.28	3.00	.097	.120
Age	20.17	0.83	22.33	4.19	3.09	.093	.123
Self-ratings[a]	4.04	1.20	4.90	1.44	2.50	.128	.102
Listening	3.58	1.51	5.08	1.78	4.96	.036	.184
Speaking	3.92	1.31	4.00	1.81	0.02	.898	.001
Reading	4.42	1.44	5.67	1.30	4.96	.037	.184
Writing	4.25	1.48	4.83	1.47	0.94	.344	.041

Note. *SD* = standard deviation. [a]Ratings scored on a 10-point scale with 1 = minimum proficiency and 10 = near-native proficiency.

Effect of proficiency of learners when using natural human speech. Table 5 shows the mean RTs of the repetition task when using natural human speech. A mixed design 2 (Proficiency) × 2 (Task) × 2 (Repetition) ANOVA was conducted with Proficiency (the upper- or lower-proficiency) as a between-subject factor, while Task (rhyme judgment task or synonym judgment task in the study phase) and Repetition (whether the vocabulary was heard once before during the study phase) as within-subject factors. Using recorded natural human speech, the results showed a significant main effect of Proficiency, F (1, 22) = 5.47, p = .03, η_p^2 = .20, and a significant interaction between Task and Repetition, F (1, 22) = 19.42, p < .001, η_p^2 = .47. The results of the simple main effect test showed a significant main effect only for when the rhyme judgment task was conducted during the study phase, F (1, 22) = 52.68, p

< .001, η_p^2 = .71, and not for when the synonym judgment task was conducted during the study phase, F (1, 22) = 2.27, p = .15, η_p^2 = .09. Figure 2 shows that though there was a perceptual learning effect if participants focused the perceptual dimension of words, the same effect was not present if they focused on the meaning of the words, regardless of proficiency when the tasks used natural human speech.

Table 5

Mean RTs (ms) of the Repetition Task of Two Proficiency Groups When Using Natural Human Speech

	Lower (n = 12)				Upper (n = 12)			
	Rhyme		Synonym		Rhyme		Synonym	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Unrepeated	837.09	93.40	862.22	66.85	777.74	92.54	762.13	103.45
Repeated	810.79	87.62	859.54	70.04	741.95	100.75	751.49	118.44

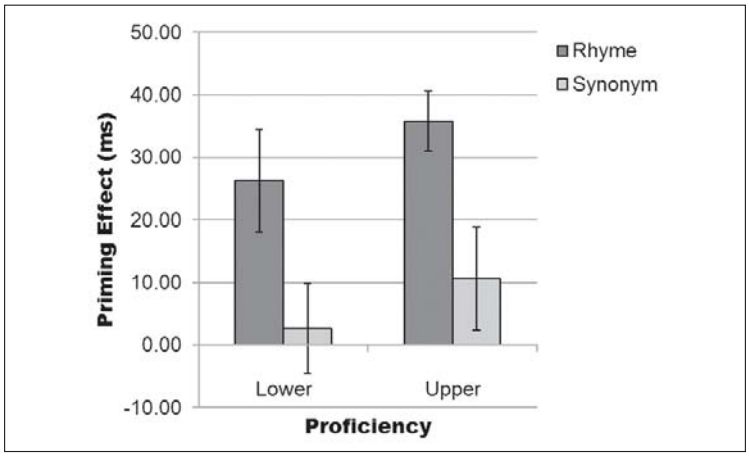

Figure 2. Mean priming effects (ms) of two proficiency groups when using natural human speech. The vertical lines indicate standard error.

Effect of proficiency of learners when using synthetic speech. Tables 6 shows the average RT of the repetition task when using synthetic speech. The same three-way ANOVA was conducted and the results showed significant main effects for Proficiency, F (1, 22) = 22.90, p < .001, η_p^2 = .51, and Repetition, F (1, 22) = 38.25, p < .001, η_p^2 = .61.

To verify hypotheses 2 and 3, a mixed 2 (Task) × 2 (Repetition) ANOVA was conducted on the data of the lower and upper proficiency groups. The results of the lower proficiency group showed a significant main effect of Repetition, F (1, 11) = 24.58, p < .001, η_p^2 = .69. The results of the upper group showed a significant interaction between Task and Repetition, F (1, 11) = 6.74, p = .02, η_p^2 = .38. The results of the simple main effect test showed a repetition effect for the synonym judgment tasks during the study phase, F (1, 11) = 15.72, p = .009, η_p^2 = .59, but not for the rhyme judgment tasks during the study phase, F (1, 11) = 2.23, p = .33, η_p^2 = .17.

The rhyme judgment task was meant to facilitate perceptual learning because the task

required attention to a perceptual dimension of words. However, the statistical results and Figure 3 revealed a smaller perceptual learning effect for upper proficiency learners compared to that of lower proficiency learners. Contrarily, for the synonym judgment, which required attention to a higher-level dimension, a semantic dimension of words, a learning effect was apparent in the data for both upper and lower proficiency learners. However, the priming effect can be seen when proficiency is low, regardless of the task in the study phase.

Table 6

Mean RTs (ms) of the Repetition Task of Two Proficiency Groups When Using Synthetic Speech

	Lower (n = 12)				Upper (n = 12)			
	Rhyme		Synonym		Rhyme		Synonym	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Unrepeated	931.50	121.54	943.88	112.07	718.02	107.15	717.39	119.36
Repeated	912.49	107.17	928.98	125.82	710.98	109.27	688.82	124.31

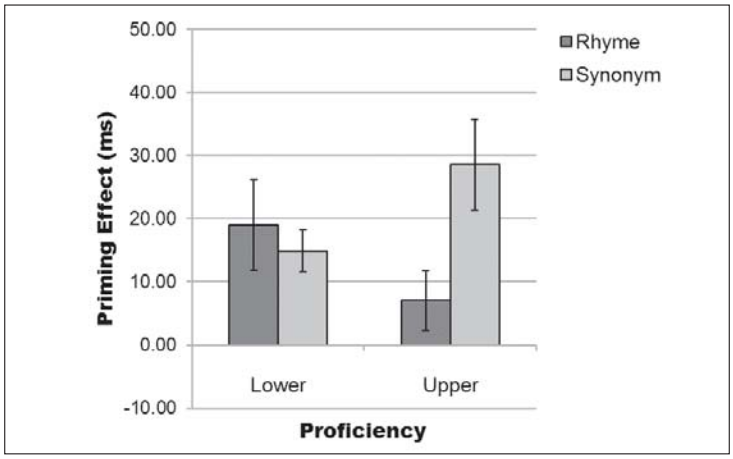

Figure 3. Mean priming effects (ms) of two proficiency groups when using synthetic speech. The vertical lines indicate standard error.

Discussion

The author believes that if synthesizing TTS speech software is effective in perceptual learning training, it can be used to freely create speech to raise the proficiency of listening skills of Japanese students learning English as a foreign language. The researcher therefore examined how the auditory learning effect from synthesized TTS differs from that of natural human speech based on three hypotheses in line with three research questions.

Natural Human Speech Versus Synthetic Speech

There is a high possibility that the results of the priming effect from natural human speech and synthetic speech support hypothesis 1 (the perceptual learning effect will be greater with natural human speech than with synthetic speech). When participants focused more on the

perceptual dimension, the results coincided with the results of previous studies on intelligibility and comprehensibility of synthetic speech processed in L1 or L2. In other words, hypothesis 1 is supported only when participants focused more on the perceptual dimension. Therefore, task differences in the study phase should be considered regarding hypothesis 2.

Task Differences in the Study Phase

Next, this study verified that hypothesis 2 (the perceptual learning effect will decrease when using natural human speech and focusing on meaning) is sound, which is in accordance with previous studies. This result confirms the complexity of the components of human speech. In addition, since this effect is unrelated to proficiency levels, it suggests that even Japanese learners of English with a higher language proficiency can have a processing delay in the perceptual stages of learning when communicating in real time, which requires a focus on meaning. This effect is even more pronounced in learners with lower language proficiency, suggesting a greater necessity for perceptual learning training.

The Effect of Proficiency Levels on Research Outcomes

Finally, considering the proficiency-based analyses, as in hypothesis 3 (the perceptual learning effect will be greater using natural human speech, but decrease using synthetic speech for learners with high proficiency levels), this study confirmed that the perceptual learning effect with synthetic speech was lower in learners with higher proficiency when focusing on the sound, but higher when focusing on the meaning.

At first glance, this may seem to contradict previous L2 priming studies. However, previous priming studies used natural human speech (except for Matsuda, 2013), whereas this study also includes synthetic speech. Since the components of synthetic speech are more controlled, such as steady reading speed and regular segmentation as previously mentioned, there are some unnatural features in the speech. Because of the unnatural features, the perceptual learning effects of synthetic speech when focusing on sound might be diminished for learners with an upper proficiency level. If it were difficult for them to become accustomed to the synthetic speech because of the unnaturalness, the mean RTs of the repetition task when using synthetic speech should be longer than the mean RTs of the repetition task with natural human speech (see the mean RTs of Table 5 & 6). However, the data suggest that they easily become accustomed to synthetic speech, in contrast with lower proficiency level learners, and respond faster to the task. If the task is too easy, there will be no priming effect because there is no learning gain,[3] as we saw in this study.

On the other hand, we can see the perceptual learning effect when focusing on meaning. The logical inference is that the semantic process facilitates the perceptual learning of students with upper proficiency levels if they do not have any problems with the stages of phonetic perception (i.e., if they can easily capture the sound of the word).

Interestingly, the priming data from the lower-proficiency group reveal that some sort of perceptual learning was facilitated regardless of the tasks in the study phase. Considering the trends revealed in previous studies that low proficiency learners prefer synthetic speech, this result suggests that using synthetic speech to facilitate perceptual learning for learners in the early stages of learning may be more successful. In light of the results of this study, synthetic

speech may be able to play a larger role in environments where there are fewer chances of encountering L2 naturally outside of the classroom as long as the purpose, proficiency, and task combinations are taken into consideration.

Though further evidence is necessary, these results suggest the need to incorporate synthesized TTS into second-language learning while taking into consideration that the perceptual learning effect will vary based on the proficiency level of the students.

Conclusion

This study investigated auditory word priming to demonstrate the applicability of using synthetic TTS in perceptual learning training. The results of this study suggest that applying synthetic TTS to perceptual learning training may be effective among Japanese students with lower proficiency levels in English provided they are placed in an environment where there are few daily interactions with L2. The lower proficiency learners demonstrated a higher possibility that synthetic TTS could be a positive contribution to perceptual learning. In the future, it should be possible to create a rich learning environment for better language acquisition by incorporating this technique into training, such as shadowing.

Furthermore, synthesized TTS shows great potential for contributing to research activities that require experiments using L2 speech to control various components. In such cases, it will be possible to conduct experiments more precisely by considering that the size of the perceptual priming effect differs based on the learner's proficiency.

Finally, the challenges of this study should be addressed. Since this study used only one version of speech synthesizing software, there is a need for future research using a number of different software packages. Furthermore, the researcher recognizes that English language classrooms often have non-native as well as native teachers; however, to this date there have been no auditory priming studies comparing native and non-native speech. This is something the researcher hopes to address further in the future. In addition, since this study was an auditory word priming experiment to verify spoken word recognition, the author is of the opinion that adding evidence in phrasal or sentence units may increase the potency of these results.

There is a high potential for using synthetic TTS to contribute to building a rich learning environment in English language classrooms, as well as outside the classrooms, in the future. If synthetic TTS is incorporated into English language learning where the level of language proficiency is taken into consideration when using it, then synthetic TTS will likely become an exceptionally convenient and effective tool.

Notes

[1] Perceptual learning effects can be defined as the changes in perceptual (or sensory) systems, as observed through behavior, such as fast and accurate recognition of the target word.

[2] The mathematic tasks were conducted as a distractor task to erase the short-term memory of participants following previous studies (Trofimovich, 2005, 2008; Trofimovich & Gatbonton, 2006).

[3] In Matsuda (2013), native English speakers showed smaller priming effects compared to

that of Japanese EFL learners.

Acknowledgements

This research was supported by a scholarship from Kwansei Gakuin University. The author is grateful to the faculty of the Graduate School of Language, Communication, and Culture of Kwansei Gakuin University. She would also like to thank Professor Katsumi Yamamoto from the University of Marketing and Distribution Sciences for his valuable comments on the experiment. Her sincere appreciation also goes to the three anonymous reviewers for their constructive comments on an earlier version of this manuscript, as well as all the participants in the study.

References

Allan, D. (2004). *Oxford placement test 2.* Oxford, England: Oxford University Press.

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.

Boersma, P., & Weenink, D. (2012). Praat: Doing phonetics by computer (Version 5.3.32) [Computer program]. Retrieved from http://www.praat.org/

Drager, K. D. R., Reichle, J., & Pinkoski, C. (2010). Synthesized speech output and children: A scoping review. *American Journal of Speech-Language Pathology, 19*, 259–273.

Globalvoice English Professional (Version 2.0.1) [Computer software]. Tokyo, Japan: Hoya.

Hirai, A., & O'ki, T. (2011). Comprehensibility and naturalness of text-to-speech synthetic materials for EFL listeners. *JACET Journal, 53*, 1–17.

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.

Kataoka, H., & Ito, M. (2013). A comparative study on reading aloud: Instructions by text-to-speech synthesis sounds and a high school Japanese English teacher. *The JACEC Bulletin, 22*(1), 39–54.

Kadota, S. (2007). *Shadoingu to ondoku no kagaku* [The science of shadowing and oral reading]. Tokyo, Japan: Cosmopier.

Kadota, S. (2015). *Shadoingu ondoku to eigo komunikeshon* [The science of shadowing, oral reading and English communication]. Tokyo, Japan: Cosmopier.

Kashiwagi, H., Kang, M., & Ohtsuki, K. (2008). Current status and future prospects of application of synthetic speech in foreign language learning, *Journal of the School of Language and Communication Kobe University, 5*, 10–19.

Koul, R., & Clapsaddle, K. C. (2006). Effects of repeated listening experiences on the perception of synthetic speech by individuals with mild-to-moderate intellectual disabilities. *Augmentative and Alternative Communication, 22*, 112–122.

Matsuda, N. (2013). Second-language speech processing: Auditory word priming in Japanese EFL learners and native English speakers, *Journal of the Japan Society for Speech Sciences, 14*, 43–62.

McDonough, K., & Trofimovich, P. (2009). *Using priming methods in second language research*. New York, NY: Routledge.

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

Oxford University Press., University of Cambridge., & Association of Language Testers in Europe. (2001). *Quick placement test*. Oxford, England: Oxford University Press.

Pinkoski-Ball, C., Reichle, J., & Munson, B. (2012). Synthesized speech intelligibility and early preschool-age children: comparing accuracy for single-word repetition with repeated exposure. *American Journal of Speech-Language Pathology, 21*, 293–301.

Reynolds, M., & Jefferson, L. (1999). Natural and synthetic speech comprehension: Comparison of children from two age groups. *Augmentative and Alternative Communication, 15*, 174–182.

Sugiura, K., & Hori, T. (2012). Auditory priming effect in Japanese learners of English: An investigation using a repetition task of spoken words. *KATE Journal, 26*, 39–51.

Tamai, K. (2005). *Risuningu shidouhou toshiteno shadoingu no kouka ni kansuru kenkyu* [A study on the effectiveness of shadowing as an instructional method of listening]. Tokyo, Japan: Kazama Shobou.

Trofimovich, P. (2005). Spoken-word processing in native and second languages: An investigation of auditory word priming. *Applied Psycholinguistics, 26*, 479–504. doi: 10.1017.S0142716405050265

Trofimovich, P. (2008). What do second language listeners know about spoken words? Effects of experience and attention in spoken word processing. *Journal of Psycholinguistic Research, 37*, 309–329. doi: 10.1007/s10936-008-9069-z

Trofimovich, P., & Gatbonton, E. (2006). Repetition and focus on form in processing L2 Spanish words: Implications for pronunciation instruction. *The Modern Language Journal, 90*, 519–535.

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

Watanabe, S., Iwaki, T., Kaneyasu, T., & Miki, K. (2006). Corpus-based Text-to-speech and its application. *Oki Technical Review, 206*(73-2), 62–65.

Yokokawa, H. (Ed.). (2006). *Kyoiku kenkyu no tameno dainigengo detabesu: Nihonjin eigogakushusha no eitango shinmitsudo mojihen* [The data base of a second language for teaching and research: Written English word familiarity of Japanese EFL learners]. Tokyo, Japan: Kuroshioshuppan.

Yokokawa, H. (Ed.). (2009). *Kyoiku kenkyu no tameno dainigengo detabesu: Nihonjin eigogakushusha no eitango shinmitsudo onseihen* [The data base of a second language for teaching and research: Spoken English word familiarity of Japanese EFL learners]. Tokyo, Japan: Kuroshioshuppan.

Appendix A

Words for Repetition Task after Rhyme Judgment Task

Word	Voice	Familiarity (Spoken)	Familiarity (Written)	Frequency (BNC)	Syllable number	Duration (ms)
			Repeated Word			
contrast	Female	5.49	5.11	8172	2	857.78
energy	Female	5.94	5.42	13083	3	671.31
husband	Female	5.60	5.08	12263	2	634.01
influence	Female	5.21	5.23	15130	3	790.65
myself	Female	6.08	5.91	12444	2	611.64
open	Female	6.13	6.77	46095	2	566.88
remember	Female	6.79	6.07	26748	3	663.85
sentence	Female	5.54	5.89	10127	2	880.16
water	Female	6.70	6.77	35767	2	604.18
addition	Male	5.97	4.70	10664	3	693.69
character	Male	5.34	6.21	12511	3	677.27
design	Male	5.27	5.75	26375	2	708.60
feature	Male	6.37	5.28	17219	2	559.42
however	Male	6.25	5.73	60498	3	589.26
listen	Male	6.34	6.17	12080	2	618.66
present	Male	6.40	5.85	36806	2	686.23
probably	Male	6.05	5.63	27303	3	716.06
recent	Male	5.79	5.25	15812	2	648.93
Average		5.96	5.71	22172.06	2.39	676.59
			Unrepeated Word			
address	Female	5.09	6.41	11984	2	745.90
basic	Female	6.04	6.05	10860	2	604.18
career	Female	5.36	5.00	9441	2	723.50
difference	Female	5.15	6.48	19138	3	790.65
island	Female	6.45	5.85	7649	2	604.18
movement	Female	5.94	5.05	17880	2	708.60
October	Female	5.61	5.88	10600	3	775.73
party	Female	6.51	6.48	52979	2	566.88
welcome	Female	6.67	6.30	9570	2	675.59
anyone	Male	6.30	5.68	14956	3	555.22
area	Male	6.17	6.00	58449	3	589.26
couple	Male	5.64	5.54	15330	2	522.13
easy	Male	6.57	6.80	21480	2	512.68
expression	Male	5.18	4.80	8756	3	835.41
money	Male	6.72	6.61	37892	2	560.05
restaurant	Male	6.82	5.90	5100	3	709.83
Sunday	Male	6.51	6.53	10100	2	753.36
suppose	Male	5.63	5.25	14482	2	835.41
Average		6.02	5.92	18702.56	2.33	670.47

Appendix B

Words for Repetition Task after Synonym Judgment Task

Word	Voice	Familiarity (Spoken)	Familiarity (Written)	Frequency (BNC)	Syllable number	Duration (ms)
			Repeated Word			
control	Female	6.13	6.49	38281	2	673.37
decide	Female	5.10	5.31	24380	2	694.41
father	Female	6.33	6.29	23216	2	652.33
history	Female	6.11	6.13	20064	3	596.21
measure	Female	6.07	4.48	17443	2	645.31
official	Female	5.49	6.22	15931	3	638.30
paper	Female	6.71	6.50	23694	2	575.17
response	Female	5.20	5.08	14627	2	848.73
window	Female	6.78	6.18	19340	2	631.29
computer	Male	6.40	6.80	16976	3	687.40
concept	Male	5.93	5.18	9093	2	743.51
ever	Male	6.55	6.16	27195	2	512.04
exercise	Male	6.52	5.78	12721	3	960.96
image	Male	6.53	6.54	11024	2	603.23
over	Male	5.73	6.55	135170	2	568.16
police	Male	5.97	6.27	27508	2	708.44
possible	Male	5.59	6.27	34178	3	687.40
realize	Male	6.05	5.50	15575	3	813.66
Average		6.07	5.99	27023.11	2.33	680.00
			Unrepeated Word			
accident	Female	6.36	6.02	8374	3	736.50
brother	Female	5.87	6.15	11757	2	540.10
council	Female	5.78	3.07	34496	2	715.46
doctor	Female	5.62	6.38	13684	2	610.24
even	Female	5.61	6.14	90473	2	568.16
happen	Female	6.08	6.02	32075	2	610.24
relation	Female	5.82	5.17	19628	3	764.56
sample	Female	5.16	5.92	8182	2	715.46
worry	Female	6.19	6.16	9006	2	589.20
already	Male	6.24	5.73	34292	3	785.60
anything	Male	6.62	6.31	28321	3	701.43
example	Male	6.87	6.11	43402	3	694.41
important	Male	5.56	6.67	39265	3	708.44
little	Male	5.70	6.47	63383	2	512.04
machine	Male	6.09	6.37	13518	2	715.46
real	Male	5.27	6.15	22982	2	596.21
suggest	Male	5.71	5.30	28665	2	932.90
summer	Male	6.71	6.73	11563	2	603.23
Average		5.96	5.94	28503.67	2.33	672.20