January 2009

A Comparison of Natural and Synthetic Speech: With and Without Simultaneous Reading

Krista Taake
Washington University in St. Louis

WASHINGTON UNIVERSITY

Department of Psychology

A COMPARISON OF NATURAL AND SYNTHETIC SPEECH: WITH AND WITHOUT SIMULTANEOUS READING

by

Krista P. Taake

A thesis presented to the
Graduate School of Arts and Sciences
of Washington University in
partial fulfillment of the
requirements for
the degree of Master of Arts

December 2009

Saint Louis, Missouri

Table of Contents

List of Tables……..iii

Abstract…...…….2

Introduction……..3

Methods……..10

Results……15

Discussion……..19

References……..24

Appendix……30

List of Tables

Table 1. Characteristics of Passage Materials Used in the Experiment…….27

Table 2. Time (Minute:Seconds) Displayed on the Computer Monitor for Each Passage in L–only, RWL, and R–only timed conditions……....28

Table 3. Proportion Correct in Different Experimental Conditions……..29

Abstract

The present study assessed college students' ability to comprehend passage materials when input is provided in different modalities: Listening–only (listening to the text; L–only), Reading–only (reading the text silently; R–only), and Reading While Listening (simultaneously reading and listening to the text; RWL). In addition, we assessed comprehension when auditory input was provided by natural (human) and synthetic (computerized) speakers. A total of 66 participants received eight passages in three different conditions (L–only, R–only, RWL) and answered multiple–choice questions following each passage. We found comprehension was significantly poorer in the L–only as compared to the RWL and R–only conditions; however, we found no difference in comprehension in the RWL and R–only conditions. In addition, we found no differences in comprehension for natural versus synthetic stimuli in any of the conditions. Our results suggest that less cognitive effort is required by the listener for auditory encoding of discourse–length material when print is available. Findings from the study are discussed in relation to previous results comparing speech perception with natural and synthetic stimuli.

A Comparison of Natural and Synthetic Speech:

With and Without Simultaneous Reading

Human–generated speech (natural speech) contains simultaneous changes in pitch, intensity, and duration of the words and speech segments (Luce, Feustel, & Pisoni, 1983; Winters & Pisoni, 2004). Computer–generated speech (synthetic speech), on the other hand, is generally produced by a text–to–speech system that uses an algorithm to translate orthographic strings of letters into auditory speech signals and generally lacks natural variations in pitch, level, and intonation. Natural speech requires little effort when listening to the speech signal, while synthetic speech seems to require more effort when listening to the speech signal.

One possible reason why synthetic speech requires more effort than natural speech is because synthetic speech consists of minimal cues to phoneme identification. More specifically, a person who listens to synthetic speech must devote more effort to phoneme identification because the synthetic speech signal lacks natural phonetic variability (Roring, Hines, & Charness, 2007; Winters & Pisoni, 2004). Also, synthetic speech lacks redundancy in the acoustic cues to speech segments that is a hallmark of natural speech. In other words, synthetic speech lacks typical cues used to identify speech segments, which could have downstream consequences on comprehension. As a result, higher order processing (syntactic and semantic) maybe limited, and an individual may have difficulty relating individual words or sentences to the overall meaning of the text. In contrast, natural speech has redundant cues and focusing on the acoustic signal of natural speech requires little effort (Winters & Pisoni, 2004). Listening to natural speech is a very well-practiced ability, since most individuals encounter natural speech in everyday life.

Therefore, it may be more likely that the listener can comprehend the overall meaning of the text because the listener is familiar with the cues of natural speech.

Despite reports that synthetic speech requires more effort than natural speech (Jenkins & Franklin, 1982; Koul, 2003), the use of text–to–speech programs has increased dramatically during the past decade. Text–to–speech systems are becoming more popular in many educational settings. One advantage of using text–to–speech systems is that schools do not need to hire readers to record students' textbooks. Instead textbooks can be scanned into the computer program and the text can easily be converted to speech; there is no human speaker who must read and record the text, which saves time, energy, and money. Consequently, many schools are using text–to–speech computer programs to record textbooks for the disabled. Also, nondisabled students have the opportunity to buy textbooks in an Etextbook format, which allows them to convert their textbook to an audio format using a text–to–speech program. While text–to–speech programs are becoming more popular, there is little applied or theoretical research that validates their use in the classroom.

Intelligibility of Natural and Synthetic Speech

Most previous work on synthetic speech has focused on differences in intelligibility between human and computer–generated speech, with the general finding that synthetic speech is found to be less intelligible than natural productions (Winters & Pisoni, 2004). Even when synthetic and natural speakers produce similar levels of intelligibility (accuracy), synthetic productions require greater effort to encode, as indexed by longer latencies required for identification (Winters & Pisoni, 2004). The Modified Rhyme Test (MRT) was developed to assess intelligibility, and the MRT has

been used to compare the intelligibility of natural and synthetic speech (Paris, Thomas, Gilson, & Kincaid, 2000). The MRT is a single–word test and requires listeners to correctly identify the word they heard from other words that deviate by a single phonetic feature (e.g., listener must differentiate *game* from *came*, *name*, *same*, *fame,* and *tame*) (Paris et al., 2000; Winters & Pisoni, 2004). Nye and Gaitenby (1973) used the MRT to compare the intelligibility of natural and synthetic stimuli and found error rates of approximately 7.6% and 2.7% for synthetic and natural productions, respectively.

Comprehension of Natural and Synthetic Speech

Although synthetic speech has been found to be less intelligible than natural speech, studies have consistently found no difference in a person's ability to comprehend natural and synthetic speech (Delogue, Conte, Sementina, 1998; Nye et al., 1975; Pisoni & Hunnicut, 1980). One reason for this dissociation between intelligibility and comprehension of natural and synthetic speech is that different types of processing are engaged in intelligibility and comprehension tasks. Intelligibility task require more bottom–up processes, while comprehension tasks require more top–down processes. More specifically, intelligibility tasks require the listener to recognize the stimuli presented, while comprehension tasks requires the listener to perform higher–level processing by extracting the underlying meaning from the acoustic signals of speech (Papadopoulos, Argyropoulos, & Kouroupetroglou, 2008). Listeners engaged in a comprehension task are able to use semantic and syntactic information in order to understand the synthetic speech signal.

Comprehension of natural and synthetic speech has generally been reported as equivalent, regardless of the quality of the synthesizer (Jenkins & Franklin, 1982).

Jenkins & Franklin (1982) examined college students' comprehension of passages when listening to short passages read by synthetic speech (Votrax VS 6.0) that was enhanced to sound more natural (e.g., changes were made to the pitch, stress and timing of the speech) to other examples of non–enhanced synthetic speech (i.e., with no changes to the acoustic features of the synthetic speech). Memory of passage materials was assessed using a free–recall test. The results indicated that the enhanced synthetic speech provided no benefit in comprehension as compared to the synthetic speech where enhancements were not introduced. Therefore, prosody is not a critical component in the recall of passage materials.

As indicated previously, research studies have investigated comprehension differences between natural and synthetic speech. These studies have consistently found no difference in a person's ability to comprehend passages read by a natural and synthetic voice (Delogue, et al., 1998; Nye et al., 1975). The studies primarily examined differences when participants were only listening to passage materials (L–only). However, Pisoni and Hunnicutt (1980) compared college students' comprehension of synthetic (MITalk text–to–speech system) and natural speech when participants listened to passage materials (L–only) versus when participants read the text silently (R–only). Comprehension of the reading materials was assessed through multiple–choice questions, and the multiple–choice test was always presented in a written format. The results indicated that participants responded more accurately on the multiple choice test when the passages were read silently (R–only) as compared to when participants heard the passage presented in an auditory format (L–only). No difference in accuracy on the

multiple choice task was found when passages were read by natural and synthetic speakers.

Although many studies (Delogue, et al., 1998; Nye et al., 1975; Pisoni et al., 1980) have examined comprehension of synthetic speech when college students' must listen to the text (L–only), no study to our knowledge has examined college students' comprehension of synthetic speech when participants simultaneously read and listen to the text (RWL). Furthermore, no study to our knowledge has compared college students' comprehension when auditory input is provided by natural and synthetic speakers when RWL.

Studies investigating comprehension of natural speech have found contradictory results regarding whether RWL improves performance over L–only or R–only. Moreno and Mayer (2002) examined college students' comprehension when participants were reading and listening to a natural speaker (RWL) and when they only listened to the passage (L–only). In the RWL condition, 19 participants heard and read a passage on how lightening is produced. In the L–only condition, another 19 participants received a verbal explanation on the same material. After participants were presented the passage, they completed a retention test, a matching test, and a transfer test to measure comprehension on the passage (refer to Moreno and Mayer, 2002 for more details). On all three tests of comprehension, performance was better when participants received bimodal input (RWL) as compared to unimodal input (L–only). The results of this study support an advantage for RWL over L–only condition; however, the study did not investigate comprehension when participants read the passage materials silently (R–only). Therefore, it is unclear if comprehension is always better when bimodal input is

received (RWL) as compared to unimodal input (R–only; L–only). One goal of the present study is to examine comprehension differences in L–only, R–only, and RWL conditions. We hope to further understand if bimodal input (RWL) improves performance as compared to unimodal input for both the R–only and L–only conditions.

Moreno & Mayer (2002) results conflict with findings from other studies (Dowell & Shmueli, 2008) that do not support the advantage of bimodal input as compared to unimodal input. For example, Dowell and Shmueli (2008) examined college students' comprehension of short email messages (very short sentences) by comparing performance for R–only, L–only, and RWL. The auditory input in the L–only and RWL conditions was provided by a natural speaker. The dependent measure was the accuracy of comprehending the written materials, which required a "yes" or "no" response. Dowell and Shmueli (2008) found no differences in accuracy when information was provided by the combined auditory visual modality (e.g., RWL) as compared to only the visual modality (R–only). However, they found that comprehension was worse when information was provided from the auditory modality (L–only) as compared to when information was provided by the auditory and visual modalities (RWL) or the visual modality (R–only). These findings then suggest that there is an inherent modality advantage in comprehension for when text is present (i.e., RWL, R-only).

One goal of the present study is to systematically compare college students' comprehension when reading silently (R–only), listening to the text (L–only) and simultaneously reading and listening to the text (RWL). Much of the literature regarding comprehension of natural speech has found inconsistent results as to whether or not RWL improves performance over R–only or L–only. Moreno and Mayer (2002) examined

comprehension in the L–only and RWL modalities; they claim that comprehension is better for bimodal input (RWL) than unimodal input (L–only). However, they did not examine comprehension in the R–only condition. Dowell and Shmueli (2008) examined comprehension in all three modalities (R–only, L–only, RWL). They found comprehension was equivalent when RWL and R–only. Also, they found comprehension was worse when L–only. We expect to find a similar patter of results as Dowell and Shmueli (2008), since our study also examines comprehension in all three modalities; we expect comprehension to be better when participants read while they listen to the text (RWL) and when they read silently (R–only) than when participants only listen to the text (L–only).

The present study will also provide important theoretical and applied contributions regarding listeners' ability to comprehend discourse materials when input is both bimodal (RWL) and unimodal (L–only; R–only) for both natural and synthetic speech. Most of the research investigating comprehension of natural and synthetic speech has only compared performance when listening to passage materials (L–only). Previous research seems to suggest that there is no difference in the comprehension of natural and synthetic speech when only listening to the text (L–only). Therefore, we also expect comprehension of natural and synthetic speech to be equivalent when participants simultaneously read and listen to passage materials (RWL). At present, ours is the only study to our knowledge that has compared comprehension in college students when auditory input is provided by both natural and synthetic speakers in the RWL and L–only conditions.

Method

Participants

A total of 68 undergraduates between the ages of 18 and 25 at Washington University in St. Louis took part in the study. Two participants' scores were not included, since they did not follow instructions. Participants received course credit for participating in the experiment. Before the experiment, all participants were asked if they were Native English speakers without any visual, hearing, or reading impairments. Only participants who stated that they were native English speakers without any visual, hearing, or reading impairments are included in the data set. The group had a mean Wechsler Adult Intelligence Scale (WAIS) vocabulary score of 15.5 (*SD* = 1.97) (Wechsler, 1997). Total testing time for each participant was approximately 1.5 hours.

Stimuli

A total of eight passages were selected from the Multi–Media Comprehension Battery (MMCB; Gernsbacher & Varner, 1988), and Listening Comprehension for Lectures, Interviews, and spoken Narratives (LISN; Tye–Murray, Sommers, Spehar, Myerson, Hale, & Rose, in press) in order to assess discourse comprehension. Passages were selected from two different batteries because neither battery had enough passages to assess all the conditions in the experiment. Also, the two comprehension measures increased the variety of materials presented to the participants. Therefore, the passage materials have more ecological validity. The LISN and MMCB were chosen over other comprehension measures because both of these measures avoided ceiling and floor performance. In addition, the LISN and MMCB are among only a few assessment

instruments designed specifically to measure spoken comprehension. More specific details about passages in each comprehension battery can be found in Table 1.

The Multi–Media Comprehension Battery (MMCB; Gernsbacher & Varner, 1988) assesses general comprehension skills. The MMCB includes three parts: auditory, written and pictorial materials. Auditory and written materials were the only materials used in the current study. Auditory and written materials consisted of four narrative passages. Following each narrative passage, 12 multiple choice questions were presented in order to test the participants' comprehension of the passage. Each multiple choice question had four possible responses. In total, there were 48 MMCB questions (12 questions for each of the four MMCB passages). All questions were presented in a written format on the computer monitor.

The Listening Comprehension for Lectures, Interviews, and Spoken Narratives (LISN; Tye–Murray, Sommers, Spehar, Myerson, Hale, & Rose, in press) assesses spoken discourse comprehension. Internal reliability as assessed with Cronbach's alpha was .75 for 18–25 year olds (Sommers, Hale, Myerson, & Rose, in press). The LISN includes six passages, including two narratives, interviews, and lectures. Narratives were acquired from Rutgers University Oral History project, in which individuals related specific life events. Lectures were acquired from the BBC Reith lectures, in which issues of contemporary interest are discussed. Interviews were selected from versions of C–span Booknotes, in which authors share information regarding their books, their research, and their lives. Only lectures and narratives were used in the current study. Interviews were excluded because the passages had two speakers, which was different from the other passages used in the experiment.

Following each passage of the LISN, six multiple choice questions are presented on the computer screen in order to test the participants' comprehension of the passage. Two questions of each type (information, integration, inference) were asked in order to assess various types of spoken discourse comprehension. *Information* questions assessed how well participants were able to recall specific details from the passage. *Integration* required participants to combine two or more pieces of information presented separately in the passages. *Inference* questions assessed how well participants were able to derive implications about information that was not explicitly stated in the passage. Each multiple choice question had four possible responses. In total, there were 24 LISN questions (6 questions for each of the four LISN passages). All questions were presented in a written format.

A survey was administered to ten disability resource centers across the country in order to select a text–to–speech reading program that is widely used in the academic community. All academic institutions surveyed used the Kurzweil 3000™ text–to–speech reading program. The Kurzweil 3000™ program consists of four male and four female synthetic voices. One male and female voice (i.e., VW Kate and VW Paul) were judged by the experimenter to be most similar to human speech; these voices were selected from the software to be used in the study. The synthetic audio recordings (16–bit with a sampling frequency of 44.1 kHz) were obtained from the software program using the text to speech synthesis algorithm in the Kurzweil 3000™ program.

The audio recordings of natural speech consisted of human recorded speech, in which three males and three females were asked to read the materials in a "natural" or "conversational" style. We used a variety of speakers because we know that difference in

intelligibility exists across speakers. Audio recordings were obtained using a XX microphone and were converted to digital files using a XX A/D converter (16–bit, 44.1kHz sampling rate). The level of all recordings were normalized to minimize level differences between the recordings. Stimulus duration and speaking rates for all stimuli can be found in Table 1. In the L–only conditions, participants looked at a blank white screen while simultaneously hearing the stories presented over headphones. The corresponding auditory materials were presented through Sony Dynamic Stereo Professional (MDR–7506) headphones at approximately 70 dB SPL.

Design

The 2 × 2 within–subjects design had 2 levels of auditory input (natural, synthetic) and 2 levels of visual input (text, no text). Passages were presented to participants in the following conditions: L–only, RWL, R–only. In the L-only condition, participants listened to a synthetic or natural voice read them passages. In the RWL condition, participants listened to a synthetic or natural voice read them passages and were simultaneously provided with the text of the passage on the computer screen. In the R–only condition participants read the text silently without receiving auditory input (R–only). There were two levels in the R–only condition (timed, self–paced). In the R–only self–paced condition, participants received as much time as they needed to read the passage silently. In the R–only timed condition, the time allotted for participants to read each passage was identical to the total spoken duration of the passage (see Table 2).

We had a total of six conditions and eight passages. The eight passages were counterbalanced across six conditions. Therefore, in each version of the experiment, there were two conditions that were repeated (Appendix A). Across all participants, all

passages and conditions were presented an equal number of times. Passages were presented in a random order, and a multiple choice test followed each passage. The multiple–choice items were in a four–alternative format, and the dependent variable was the proportion correct responses on the multiple choice tests.

Procedure

Participants were provided with verbal and written instructions for the experiment. During the first part of the experiment, participants sat in front of a computer monitor wearing headphones. Short stories were presented orthographically on a 17–in. computer monitor in SuperLab 4.0. Participants read the passages on a computer screen and/or heard the passages presented through the headphones. In the R–only self–paced condition, participants could spend as long as they wanted with each portion of the passage on the screen and pressed a key to advance to the next screen when they were ready for subsequent portions of the passage. In the R–only timed and RWL conditions, approximately two to three paragraphs of each passage were displayed on the screen for a designated amount of time. The screen automatically advanced to the next set of paragraphs after the designated amount of time, which was the same value in the R–only timed and RWL conditions for each passage (Table 2).

Immediately following each passage, participants answered multiple choice questions, which were presented on the computer screen. The participants responded to the question by pressing one of four response buttons on the keyboard. Participants had an unlimited amount of time to answer each question but were not allowed to go back to the text or to replay any of the auditory passages. Feedback was not provided. In order to prevent fatigue, participants were instructed to take a 15 minute break after they finished

answering the questions for the fourth passage. After the 15 minute break, participants finished reading and answering questions about the remaining four passages.

After answering questions for all the passages, participants were asked to complete a questionnaire stating their opinion of their experience listening to the synthetic/natural voices. The questionnaire was adopted from a mean opinion score (MOS) that measures speech quality for text–to–speech systems (Viswanathan & Viswanathan, 2004). Participants were asked to state their opinions of the overall sound quality of the natural and synthetic voices. The questionnaire had 12 sections: Overall impression; Listening effort; Pronunciation; Speaking rate; Voice pleasantness; Voice naturalness; Audio flow; Ease of listening; Comprehension; Articulation; Performance; Acceptance. All sections except one were presented in a randomized order; questions within each section were also randomized. The section assessing acceptance was presented last in order to prevent biases to other sections (‘‘Do you think that the computer voice can be used as an alternative to books on tape for the reading/visually impaired?’’). Items from the questionnaire can be found in Appendix B. Following the questionnaire, participants were instructed to take another 15 minute break. Participants then completed the WAIS vocabulary test.

Results

Scoring

Participants received 6 question following each LISN passage and 12 questions following each MMCB passage. We calculated the proportion of correct responses for each participant from all the questions. Scores on the multiple choice tests assessed comprehension in six conditions (R–only timed; R–only self–paced; L–only natural; L–

only synthetic; RWL natural; RWL synthetic). The total number of questions was the same in each condition. Participants received 1–point for each correct answer and 0–points for each incorrect answer. No partial credit was awarded. A summary of our results are found in Table 3. In all the analyses to follow, the WAIS vocabulary scores were used as a covariate.

Comparison of modalities: RWL, L–only, R–only timed

The data were subjected to an analysis of variance with one within–subject factor, modality, which consisted of three levels (RWL, L–only, R–only timed). We did not include R–only self–paced because participants received a different amount of time to read passages in this condition; therefore, we wanted to control for the time the passages were presented. The descriptive data are displayed in Table 3. There were differences in comprehension among the RWL, L–only, and R–only timed conditions as indicated by a significant main effect of modality , $F(2, 128) = 4.45$, $p < .05$. Pairwise post hoc analyses using a Bonferroni correction for multiple comparisons indicated that comprehension was poorer when participants listened to the text (L–only modality) as compared to when they had the text available (RWL, R–Only timed).

Furthermore, there was no difference in comprehension when participants read the text (R–only timed) and when they simultaneously read and listened to the text (RWL), indicating that there wasn't a benefit in being provided additional auditory information when text is already available.

Comparison of R–only timed and R–only self–paced

It should be noted that there were two levels in the R–only condition. In one condition, participants received as much time as they needed to read the passages (R–

only self–paced), while in the other condition participants were provided the same amount of time for the passages as presented in the L–only and RWL conditions (R–only timed). A within–subject analysis of variance was conducted on the proportion of correct responses for the two levels in the R–only condition. Comprehension was not significantly different for passages presented with unlimited reading time and those presented under timed conditions $F(1, 65) = .96$, $p > .05$ (Table 3).

Comparison of modalities and speech types

To determine if comprehension differed as a function of speech type (natural versus synthetic) and to examine any interactions between modality and speech type, we conducted an ANOVA with two within–subject factors, speech type (natural, synthetic) and modality (L–Only, RWL). Our results indicated that overall comprehension scores did not differ according to whether or not participants listened to a natural or synthetic speaker, $F(1, 64) = 1.53$, $p > .05$. However, comprehension scores did significantly differ according to the different modalities, $F(1, 64) = 4.65$, $p < .05$. Pairwise post hoc analyses using a Bonferroni correction for multiple comparisons indicated that comprehension was better when participants simultaneously read and listened to the text (RWL) as compared to when participants only listened to the text (L–Only). Of particular interest, no interaction was found between speech type and modality. The lack of an interaction indicates that comprehension in the different modalities was not influenced by speech type, $F(1, 64) = .73$, $p = .39$ (Table 3).

Comparison of passage type (MMCB Narratives, LISN Narratives, and LISN Lectures)

Different types of processing are required during the reading process for different types of passages (Kintsch & Young, 1984). For example, it is assumed that readers have

prior knowledge about the structure and convention of narrative scripts. Therefore, readers of narrative passages are able to see how causal structures of events relate to one another. In contrast, lectures lack the organization components of narratives, and lectures often require readers to comprehend a series of facts (Kintsch & Young, 1984). Therefore, we wanted to determine if differences in comprehension occurred across the different passage types used in the present study. We conducted a within–subject analysis of variance on the proportion of correct responses for the passage types, which consisted of three levels (MMCB Narratives, LISN Narratives, and LISN Lectures). Our results indicated that overall comprehension scores did not differ according to the passage type, $F(1, 128) = 1.55$, $p > .05$ (proportions correct: MMCB Narratives = .73, LISN Narratives = .73, LISN Lectures = .65). Due to limited stimuli, we were not able to determine if an interaction was present between passage type (MMCB Narratives, LISN Narratives, and LISN Lectures) and speech type (Natural, Synthetic).

Comparison of question type

Following the LISN passages, different types of questions (information, integration, and inference) were asked to assess comprehension. Each participant answered eight questions for each question type (two questions on each of the four LISN passages). A within–subject analysis of variance was conducted on the proportion of correct responses for the three types of questions. The results indicated no main effect in accuracy for the different types of questions, $F(1, 128) = 1.54$, $p > .05$ (proportions correct: Information = .80, Integration = .61, Inference = .66). We were not able to determine if an interaction exists between question type (information, integration, and inference) and speech type (natural, synthetic).

Discussion

The present study had two main goals: (a) to compare comprehension in R–only, L–only, and RWL conditions; and (b) to compare comprehension when auditory input is provided by natural and synthetic productions. In addressing our first goal (a), we found that comprehension was significantly poorer when participants only listened to the passage (L–only) as compared to when they either simultaneously read and listened to the passage (RWL) or read the text silently (R–only); however, we found no difference in comprehension when participants simultaneously read and listened to the passage (RWL) and read the text silently (R–only). In addressing our second goal (b), we found that there was no difference in comprehension for natural versus synthetic auditory productions.

Many previous studies (Delogue, et al., 1998; Nye, et al., 1975) comparing comprehension of natural and synthetic speech have only examined performance when listening to the text (L–only). These studies consistently found that comprehension in the L–only condition is equivalent for natural and synthetic speech (Delogue, et al., 1998; Nye, et al., 1975). To our knowledge, the present study is the first study to have examined comprehension of natural and synthetic speech in both the L–only and RWL conditions in college students. We replicated the findings of Delogue, et al. (1998) and Nye, et al. (1975) for L–only, and we also found no difference in the comprehension of natural and synthetic speech in the RWL condition. Our results suggest that comprehension is similar for natural and synthetic speech; however, the level of comprehension depends on the modality of auditory input. Comprehension for natural and synthetic speech was worse for L–only as compared to RWL presentations.

There is evidence that intelligibility is poorer for synthetic speech than natural speech (Nye et al., 1973; Paris et al., 2000). In contrast, performance for natural and synthetic speech in

comprehension tasks has generally been found not to differ (Delogue, et al., 1998; Nye, et al., 1975). The dissimilar results for measures of intelligibility and comprehension tasks may be attributed to the fact that listeners are using more bottom–up processes for measures of intelligibility and more top–down processes for comprehension tasks. For example, in intelligibility tasks, participants are often required to identify words. In contrast, in comprehension tasks, participants have more time for accessing word meanings and making inferences (Koul, 2003). Also, participants would have the opportunity to use other information presented to help them complete the task (e.g., semantic, syntactic information). It is easier for participants to extract the underlying meaning from the synthetic speech signal for comprehension tasks than intelligibility tasks. Therefore, comprehension may not require perfect intelligibility.

The results of the current study are consistent with previous research that has examined college students' comprehension of discourse–length material in different modalities (Dowell & Shmueli, 2008; Moreno & Mayer, 2002). Dowell and Shmueli (2008) examined comprehension of short email messages using R–only, L–only, and RWL presentations, and they found comprehension was significantly poorer when participants only listened to the passage (L–only) as compared to when they simultaneously read and listened to the passage (RWL) and read the text silently (R–only); similar to the present results, they also found no difference in comprehension when participants simultaneously read and listened to the passage (RWL) and read the text silently (R–only). It should be noted that Dowell and Shmueli (2008) examined comprehension when auditory input was provided by a natural speaker. However, as noted above, the present study found the same pattern of results when auditory input was provided by a synthetic speaker for the L–only and RWL conditions.

Moreno and Mayer (2002) examined college students' comprehension when participants read and listened to the text (RWL) and listened to the text (L–only). Auditory input was provided by a natural speaker. They found performance was better when participants simultaneously read and listened to the text (RWL) than when they only listened to the text (L–only). Moreno and Mayer (2002) results are also similar to the current study. However, Moreno and Mayer (2002) interpreted their results differently than us.

Specifically, Moreno and Mayer (2002) argued that bimodal input improves performance; however, as noted previously, our results and Dowell and Shmueli (2008) do not support an advantage for bimodal input over all types of unimodal input (e.g., R–only). There appear to be mixed conclusions as to whether RWL performance improves comprehension. Unlike Moreno and Mayer (2002), our study and Dowell and Shmueli (2008) examined performance when participants read the text silently (R–only). Dowell and Shmueli (2008) as well as the current study found that performance was the same in the RWL condition and R–only condition. Therefore, RWL does not improve comprehension over R–only; performance is the same (Table 3).

The discrepancy between our conclusion and Moreno and Mayer (2002) could also be attributed to different types of testing and passage materials. We tested participants' comprehension on LISN and MMCB passages by using a multiple choice exam. Moreno and Mayer (2002) tested participants' comprehension on a passage describing how lightning is produced through a retention test, matching test, and transfer test (see Moreno & Mayer, 2002 for more details). However, Dowell and Shmueli (2008) used different passage materials and a different type of test than the current study, and their results also do not support an advantage for RWL as compared to R–only.

In our attempt to reconcile the different interpretations, we propose that all three studies (our study, Dowell & Shmueli, 2008; Moreno & Mayer, 2002) support the idea that it is easier (less cognitively effortful) to understand written text (e.g., R–only, RWL) than heard discourse (L–only). Adding auditory information does not add significantly to comprehension performance when text is available, even when different tests and passage materials are used. Although our conclusion applies across various studies (our study, Dowell & Shmueli, 2008; Moreno & Mayer, 2002), all of the studies compared comprehension in college students where it is expected that advanced reading skills are present. However, the skill of the reader may be important when comparing comprehension differences when reading silently (R–only), listening to the text (L–only), and simultaneously reading and listening to the text (RWL).

Future research needs to be conducted to determine exactly how the skill of the reader could influence comprehension abilities in different modalities. The current study primarily tested individuals who likely have high comprehension abilities. The Wechsler Adult Intelligence Scale (WAIS) vocabulary test is a strong predictor of reading comprehension abilities. Our participants had a mean WAIS vocabulary score of 15.5 (*SD* = 1.97), indicating that the population that we tested is in the upper 95^{th} percentile in their comprehension abilities. However, it could be the case that when literacy ability of the reader is poor, participants may perform worse when only written text is available (R–only) as compared to when auditory input is available (L–only; RWL). More work in the field needs to be done to determine exactly how the skill of the reader affects comprehension in different modalities.

The results of the current study have several important implications. Presently, there has been relatively little research assessing comprehension differences across different modalities (R–only, L–only, RWL) for natural and synthetic speech. Most research in the field has only

compared comprehension of natural and synthetic speech when only listening to text (L–only). Furthermore, our study is the only study to our knowledge that has compared college students' comprehension of natural and synthetic speech in the following conditions: L–only, RWL, and R–only. The results of the current study have many important implications and have the potential to be applied to educational settings in order to help students learn more effectively. More specifically, students may benefit from computer reading programs that read textbooks to them as long as written text is present. However, as noted above, there may be other factors beyond the scope of this study that may limit comprehension. More research should be conducted before these results are applied in the classroom.

References

Delogu, C., Conte, S. & Sementina, C. (1998). Cognitive factors in the evaluation of synthetic speech. *Speech Communication, 24,* 153–168.

Dowell, J., & Shmueli, Y. (2008). Blending speech output and visual text in the multimodal interface. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 50,* 782–788.

Gernsbacher, M. A., & Varner, K. R. (1988). *The multi–media comprehension battery*. (Tech. Rep. No. 88–07). Eugene, OR: University of Oregon, Institute of Cognitive and Decision Sciences.

Hartman, F. R. (1961). Single and multiple channel communications: A review of research and a proposed model. *AV Communication Review, 9,* 235–262.

Jenkins, J. J., & Franklin, L. D. (1982). Recall of passages of synthetic speech. *Bulletin of the Psychonomic Society, 20,* 203–206.

Kintsch, W., & Young, S. R. (1984). Selective recall of decision–relevant information from texts. *Memory & Cognition, 12,* 112–117.

Koul, R. (2003). Synthetic speech perception in individuals with and without disabilities. *Augmentation and Alternative Communication, 19,* 49–58.

Luce, P. A., Feustel, T. C. & Pisoni, D. B. (1983). Capacity demands in short–term memory for synthetic and natural speech. *Human Factors, 25,* 17–32.

Montali, J., & Lewandowki, L. (1996). Bimodal reading: benefits of talking computers for average and less skilled readers. *Journal of Learning Disabilities 29,* 271–279.

Moreno, R., & Mayer, R. E. (2002). Verbal redundancy in multimedia learning: When reading helps listening. *Journal of Educational Psychology, 94,* 156–163.

Nye, P.W., & Gaitenby, J.H. (1973). Consonant intelligibility in synthetic speech and in a natural speech control (modified rhyme test results). In *Haskins Laboratories Status Report on Speech Research, SR–33,* 77–91. New Haven, CT: Haskins Laboratories.

Nye, P. W., Ingemann, F. & Donald, L. (1975). Synthetic speech comprehension: a comparison of listener performances with and preferences among different speech forms. In *Haskins Laboratories Status Report on Speech Research, SR–41,* 117–126. New Haven, CT: Haskins Laboratories.

Papadopoulos, K., Argyropoulos, V., & Kouroupetroglou, G. (2008). Discrimination and comprehension of synthetic speech by students with visual impairments: the case of similar acoustic patterns. *Journal of Visual Impairment and Blindness*, *102*, 420–429.

Paris, C. R., Thomas, M. H., Gilson, R. D., & Kincaid, J. P. (2000). Linguistic cues and memory for synthetic and natural speech. *Human Factors, 42,* 421–431.

Pisoni, D. B., & Hunnicutt, S. (1980). Perceptual evaluation of MITalk: The MIT unrestricted text–to–speech system. In *1980 IEEE International Conference on Acoustics, Speech and Signal Processing,* 572–575. New York: IEEE.

Viswanathan, M., & Viswanathan, M. (2005). Measuring speech quality for text–to–speech systems: development and assessment of a modified mean opinion score (MOS) scale. *Computer Speech and Language, 19,* 55–83.

Roring, R. W., Hines, F. G., & Charness, N. (2007). Age difference in identifying words in synthetic speech. *Human Factors, 49,* 25–31.

Sommers, M., Hale, S., Myerson, J., Rose, N., Tye–Murray, N., & Spehar, B. (in press). Spoken discourse comprehension across the adult lifespan.

Tye–Murray, N., Sommers, M., Spehar, B., Myerson, J., Hale, S., & Rose, N. (in press). Auditory–visual discourse comprehension by older and young adults in favorable and unfavorable conditions, *International Journal of Audiology.*

Winters, S. J., & Pisoni, D. B. (2004). Perception and comprehension of synthetic speech. In *Research on spoken language processing report no. 26* (pp. 95–138). Bloomington: Indiana University, Speech Research Laboratory, Department of Psychology.

Wechsler, D. (1997). *Administration and scoring manual: WAIS–III. San Antonio,* TX: Harcourt Brace & Co.

Table 1

Characteristics of Passage Materials Used in the Experiment

Passage	Comprehension Measure	Passage Type	Speaker	Total Words	Words/Min	Time Natural	Time Synthetic
1	MMCB	Narratives	Woman	879	184	4:47	4:47
2			Man	957	201	4:46	4:46
3			Woman	639	172	3:43	3:43
4			Man	539	156	3:27	3:27
5	LISN		Woman	683	187	3:39	3:39
6			Man	615	224	2:45	2:45
7		Lectures	Woman	633	157	4:02	4:02
8			Man	469	164	2:52	2:52

Standard Deviations are in parenthesis

Table 2.

Time (Minute:Seconds) Displayed on the Computer Monitor for Each Passage in L–only, RWL, and R–only timed conditions

Passage	Part 1	Part 2	Part 3	Part 4	Part 5	Part 6	Total
1– MMCB 1	0:43	0:51	0:49	1:00	0:45	0:39	4:56
2– MMCB 2	0:39	1:14	0:37	1:08	0:38	0:29	5:09
3– MMCB 3	0:42	0:56	0:37	0:55	0:33		3:57
4– MMCB 4	0:51	0:38	0:49				2:34
5–LISTN 1	0:55	1:13	0:57	0:55			4:09
6–LISTN 2	0:50	0:47	1:17				2:52
7–LISTN 3	1:03	1:06	0:35	0:56			3:51
8–LISTN 4	0:43	0:40	0:53	0:29			2:59

Participants in the R–only self–paced condition received an unlimited amount of time to read each part of the passage.

Table 3

Proportion Correct in Different Experimental Conditions

Condition	Condition	Proportion Correct
Natural	RWL	0.73 (.19)
Synthetic		0.73 (.17)
Natural	L–Only	0.66 (.22)
Synthetic		0.66 (.18)
Control Timed	R–Only	0.74 (.18)
Control Self–paced		0.73 (.17)

Standard Deviations are in parenthesis

Appendix A

Session	Passage	Condition
1	1- MMCB 1	1-Natural, Text, Time
	2- MMCB 2	2-Natural, No Text, Time
	3- MMCB 3	3-Synthetic, Text, Time
	4- MMCB 4	4-Synthetic, No Text, Time
	5-LISN 1	5-No Voice, Text, Time
	6-LISN 2	6-No Voice, Text, No Time
	7-LISN 3	1-Natural, Text, Time
	8-LISN 4	2-Natural, No Text, Time
2	1- MMCB 1	2-Natural, No Text, Time
	2- MMCB 2	3-Synthetic, Text, Time
	3- MMCB 3	4-Synthetic, No Text, Time
	4- MMCB 4	5-No Voice, Text, Time
	5-LISN 1	6-No Voice, Text, No Time
	6-LISN 2	1-Natural, Text, Time
	7-LISN 3	2-Natural, No Text, Time
	8-LISN 4	3-Synthetic, Text, Time
3	1- MMCB 1	3-Synthetic, Text, Time
	2- MMCB 2	4-Synthetic, No Text, Time
	3- MMCB 3	5-No Voice, Text, Time
	4- MMCB 4	6-No Voice, Text, No Time
	5-LISN 1	1-Natural, Text, Time
	6-LISN 2	2-Natural, No Text, Time
	7-LISN 3	3-Synthetic, Text, Time
	8-LISN 4	4-Synthetic, No Text, Time

Session	Passage	Condition
4	1- MMCB 1	4-Synthetic, No Text, Time
	2- MMCB 2	5-No Voice, Text, Time
	3- MMCB 3	6-No Voice, Text, No Time
	4- MMCB 4	1-Natural, Text, Time
	5-LISN 1	2-Natural, No Text, Time
	6-LISN 2	3-Synthetic, Text, Time
	7-LISN 3	4-Synthetic, No Text, Time
	8-LISN 4	5-No Voice, Text, Time
5	1- MMCB 1	5-No Voice, Text, Time
	2- MMCB 2	6-No Voice, Text, No Time
	3- MMCB 3	1-Natural, Text, Time
	4- MMCB 4	2-Natural, No Text, Time
	5-LISN 1	3-Synthetic, Text, Time
	6-LISN 2	4-Synthetic, No Text, Time
	7-LISN 3	5-No Voice, Text, Time
	8-LISN 4	6-No Voice, Text, No Time
6	1- MMCB 1	6-No Voice, Text, No Time
	2- MMCB 2	1-Natural, Text, Time
	3- MMCB 3	2-Natural, No Text, Time
	4- MMCB 4	3-Synthetic, Text, Time
	5-LISN 1	4-Synthetic, No Text, Time
	6-LISN 2	5-No Voice, Text, Time
	7-LISN 3	6-No Voice, Text, No Time
	8-LISN 4	1-Natural, Text, Time

Appendix B

Questionnaire assessing perceptual differences between natural and synthetic speech

Overall Impression:

1. How do you rate the quality of the audio you just heard with the *human* voice?
a. Excellent (32%)
b. Good (64%)
c. Fair (5%)
d. Poor
e. Very Poor

2. How do you rate the quality of the audio you just heard with the *computer* voice?
a. Excellent (11%)
b. Good (32%)
c. Fair (36%)
d. Poor (17%)
e. Very Poor (5%)

Listening Effort:

3. How would you describe the effort you were required to make in order to understand the passage with the *human* voice?
a. Complete relaxation possible; no effort required (24%)
b. Attention necessary; no appreciable effort required (65%)
c. Moderate effort required (11%)
d. Considerable effort required
e. No meaning understood with any feasible effort

4. How would you describe the effort you were required to make in order to understand the passage with the *computer* voice?
a. Complete relaxation possible; no effort required
b. Attention necessary; no appreciable effort required (33%)
c. Moderate effort required (42%)
d. Considerable effort required (24%)
e. No meaning understood with any feasible effort

Pronunciation:

5. Did you notice anomalies in pronunciation with the *human* voice?
a. No (51%)
b. Yes, but not annoying (31%)
c. Yes, slightly annoying (15%)
d. Yes, annoying (2%)
e. Yes, very annoying (2%)

6. Did you notice anomalies in pronunciation with the *computer* voice?
a. No (6%)
b. Yes, but not annoying (15%)
c. Yes, slightly annoying (47%)
d. Yes, annoying (21%)
e. Yes, very annoying (11%)

Speaking rate:

7. The average delivery with the *human* voice was:
a. Just right (50%)
b. Slightly fast (8%)
c. Slightly slow (36%)
d. Fairly fast (2%)
e. Fairly slow (5%)
f. Very fast
g. Very slow
h. Extremely fast
i. Extremely slow

8. The average delivery with the *computer* voice was:
a. Just right (10%)
b. Slightly fast (34%)
c. Slightly slow (28%)
d. Fairly fast (8%)
e. Fairly slow (20%)
f. Very fast
g. Very slow
h. Extremely fast
i. Extremely slow

Pleasantness:

9. In general, how would you describe the pleasantness of the *human* voices?
a. Very pleasant (21%)
b. Pleasant (52%)
c. Neutral (27%)
d. Unpleasant
e. Very unpleasant

10. How would you describe the pleasantness of the *computer* voice?
a. Very pleasant
b. Pleasant (6%)
c. Neutral (35%)
d. Unpleasant (44%)
e. Very unpleasant (15%)

Naturalness:

11. How would you rate the naturalness of the audio with the *human* voice?
a. Very Natural (38%)
b. Natural (53%)
c. Neutral (5%)
d. Unnatural (5%)
e. Very Unnatural

12. How would you rate the naturalness of the audio with the *computer* voice?
a. Very Natural
b. Natural (2%)
c. Neutral (6%)
d. Unnatural (63%)
e. Very Unnatural (29%)

Audio flow:

13. How would you describe the continuity of the flow of the audio with the *human* voice?
a. Very smooth (14%)
b. Smooth (71%)
c. Neutral (15%)
d. Discontinuous
e. Very discontinuous

14. How would you describe the continuity of the flow of the audio with the *computer* voice?
a. Very smooth
b. Smooth (6%)
c. Neutral (14%)
d. Discontinuous (65%)
e. Very discontinuous (15%)

Ease of listening:

15. Would it be easy or difficult to listen to the *human* voice for long periods of time?
a. Very easy (20%)
b. Easy (55%)
c. Neutral (21%)
d. Difficult (5%)
e. Very difficult

16. Would it be easy or difficult to listen to the *computer* voice for long periods of time?
a. Very easy
b. Easy (9%)
c. Neutral (14%)
d. Difficult (52%)
e. Very difficult (26%)

Comprehension Problems:

17. Did you find certain words hard to understand when you were listening to the *human* voice?
a. Never (47%)
b. Rarely (47%)
c. Occasionally (6%)
d. Often
e. All of the time

18. Did you find certain words hard to understand when you were listening to the *computer* voice?
a. Never (6%)
b. Rarely (26%)
c. Ocasionally (61%)
d. Often (6%)
e. All of the time (2%)

Articulation:

19. Were the sounds in the audio distinguishable when you were listening to the *human* voice?
a. Very Clear (53%)
b. Clear (44%)
c. Neutral (3%)
d. Less Clear
e. Much less Clear

20. Were the sounds in the audio distinguishable when you were listening to the *computer* voice?
a. Very Clear (6%)
b. Clear (29%)
c. Neutral (17%)
d. Less Clear (42%)
e. Much less Clear (6%)

Performance:

21. How easy or difficult was the multiple choice test after listening to the passage read by the *human* voice *without* having the text in front of you?
a. Very easy (6%)
b. Easy (39%)
c. Neutral (36%)
d. Difficult (14%)
e. Very difficult (5%)

22. How easy or difficult was the multiple choice test after listening the passage read by the *computer* voice *without* having the text in front of you?
a. Very easy (5%)
b. Easy (12%)
c. Neutral (24%)
d. Difficult (50%)
e. Very difficult (9%)

23. How easy or difficult was the multiple choice test after listening to the passage read by the *human* voice *with* having the text in front of you?
a. Very easy (26%)
b. Easy (55%)
c. Neutral (14%)
d. Difficult (3%)
e. Very difficult (2%)

24. How easy or difficult was the multiple choice test after listening the passage read by the *computer* voice *with* having the text in front of you?
a. Very easy (14%)
b. Easy (42%)
c. Neutral (23%)
d. Difficult (20%)
e. Very difficult (2%)

Acceptance:

25. Do you think that the *computer* voice can be used for as an alternative to books on tape for the reading/visually impaired?

a. Yes (16%)

b. No (84%)