The effects of time compression on the comprehension of natural and synthetic speech

Janine Lebeter & Susan Saunders
University of Victoria
jlebeter@uvic.ca

Synthetic speech is commonly used as the output signal in text-to-speech synthesizers. The purpose of this study is to determine if high quality synthetic speech, such as the type used by speech-generating devices, is perceived as well as natural human speech. Little research has looked at the comprehension of synthetic versus natural speech through the dimension of time compression. This research fills that gap by comparing the comprehension of time-compressed natural speech signals and time-compressed synthetic speech signals. A secondary aim is to determine the quality of current text-to-speech (TTS) synthesizers that come with current (as of 2010) computers. In this experiment, signal comprehension was tested with a speeded sentence verification task. Participants were able to verify natural speech sentences faster and more accurately than synthetic speech sentences. Additionally, as the sentence compression rate was increased, comprehension became more difficult for both speech conditions, with the greatest adverse affect being found for synthetic speech comprehension.

1 Introduction

Synthetic speech generators have become an important tool in the lives of many individuals. It is common for people with language disorders and delays to use speech synthesizers to augment their communication, and within the past few years, both Windows and Macintosh computers have equipped their newest models with pre-installed text-to-speech generators. The majority of these devices use text-to-speech synthesis wherein graphemes, digits, and words are entered using a keyboard or touch screen as input, which is then converted into a synthetic speech waveform by a set of algorithmic rules (Koul, 2003; Koul & Clappsaddle, 2006). Studies investigating what effect, if any, speech generating devices (SGDs) have on the lives of individuals with mild to severe intellectual disabilities, visual impairments, and special communication needs have clearly shown that SGDs make a profound difference. These devices have been shown to be more effective than vocalizations, gestures and non-electronic communication boards in conveying information, and they lead to an increase in positive communicative interactions with peers and support personnel (Koul, 2003; Koul & Clapsaddle, 2006; Koul & Hester, 2006). Even though SGDs play such an important role in the lives of many individuals, the quality of the synthetic speech is not guaranteed, and depending on the sophistication of the device, the output may vary greatly (Koul, 2003).

Since the middle to late 1980s, many researchers have compared the quality of synthetic versus natural speech (Hoover et al. 1987; Logan et al., 1989; Mitchel & Atkins, 1989; as cited in Koul, 2003; Mirenda & Beukelman, 1987). Speech signal quality is discussed in terms of intelligibility and comprehension, where intelligibility refers to an individual's ability to recognize phonemes and words presented in isolation, while comprehension requires that a listener transform the linguistic message into a meaningful mental representation (Koul, 2003). The present study tests for signal comprehension because it is important that people are able to construct meaningful mental representations from the synthetic speech used in speech generating devices. Koul (2003) suggests that, for single word identification tasks conducted in ideal listening conditions, there is no significant difference between the perception of high-quality synthetic speech and natural speech. Other research has supported the opposite view: that digitized or synthetic speech is more difficult to perceive than natural speech and demands greater cognitive resources to process (Duffy & Pisoni, 1992; Francis & Nusbaum, 2009; Mirenda & Beukelman, 1987). Since the advent of the first text-to-speech computer-based system in 1968, formant synthesis technology has greatly improved. One aim of the present study is to assess the quality of current text-to-speech synthesizers that come pre-installed in new computers.

Past research has tested the quality of synthetic speech by manipulating variables such as background noise, age of listener, intellectual ability of listener, and experience with the signal (Koul, 2003; Koul & Hanners, 1997; Mirenda & Beukelman, 1987). However, to the best of our knowledge, no study has tested the quality of synthetic speech by manipulating speech rate. For participants, fast speech rates create an adverse listening condition (Adank & Devlin, 2010; Adank & Janse, 2009; Dupoux & Green, 1997; Golomb, Peelle, & Wingfield, 2007; Janse, 2004; Pallier & Sebastian-Gallés, 1998), which is desirable when signal quality is being tested. Additionally, Dupoux and Green (1997) have pointed to time-compressed speech as being an ideal independent variable for a number of reasons. Firstly, with linear time compression, speech signals can be altered in quantifiable and measurable ways to create stimuli that are outside the bounds of everyday experience. Secondly, newer compression algorithms such as Praat's "Pitch Synchronous Overlap and Add" (PSOLA) function (Boersma & Weenik, 2009), used in the present study, allow speech to be compressed without deleting segments of the original signal or creating discontinuities, which was common with older compression techniques. Finally, compressed speech affects the perceived rate at which the signal was produced. Because of this, it has been argued that any perceptual effects found for time compressed speech can be compared with, and generalized to, more natural changes in speech rate (Dupoux & Green, 1997).

Adank and Janse's (2009) study of perceptual learning mechanisms used naturally fast and linearly time-compressed speech to study human adaptation to atypical speech signals. Participants were asked to perform a speeded sentence verification task for both naturally fast and artificially time-compressed stimuli. Surprisingly, the researchers found that time-compressed speech was easier to adapt to – as measured by faster reaction times and overall higher verification accuracy – than natural fast speech. This finding supports past research, which had found that natural fast speech is difficult to adapt to because it is not only temporally compressed, but it is also spectrally different from regular conversational speech (Janse, 2004; Adank & Janse, 2009). The spectral variation that occurs with fast speaking rates is caused by the increased occurrence of coarticulation and segment deletion, a change in the overall prosodic pattern of the speech stream, and a tendency to reduce the duration of vowels and unstressed syllables.

The naturally fast stimuli used in Adank and Janse's (2009) study were all produced by a single individual. The speaker was instructed to read 180 experimental sentences aloud as declarative statements at his normal speaking rate, while recordings were taken. He was then instructed to produce all of the stimuli again by reading each individual sentence aloud four times in quick succession so as to achieve a very fluent speaking rate. It was found that, on average, the fast versions of the sentences were compressed to approximately 46% of the original sentence duration, with the fastest items being produced at approximately 33% of the original sentence duration (Adank & Janse, 2009). Given that such fast speech rates are achievable by human articulation, we predicted that our participants would be able to comprehend at least some sentences presented at such fast rates, as they will have had experience with these fast speech rates during their lifetime. Dupoux and Green (1997) also analyzed perceptual adjustment mechanisms for highly compressed natural speech. Their fast stimuli were compressed to 38% and 45% of the original speaking rate. It was found that the sentences compressed to 38% of their original duration were difficult for participants to understand, and that the adjustment process took longer for stimuli that had been compressed to a greater degree (Dupoux & Green, 1997). The literature shows that increased speech rates are more difficult to perceive and comprehend than conversational speech rates (Adank & Devlin, 2010; Adank & Janse, 2009; Dupoux & Green, 1997; Golomb, Peelle, & Wingfield, 2007; Janse, 2004; Pallier & Sebastian-Gallés, 1998). For the present study, it is predicted that, as the compression rate is increased across experimental blocks, signal comprehension will become more difficult in both the natural and synthetic stimuli conditions. The speech signal that facilitates comprehension at a higher rate of compression will be considered to be the higher quality signal because it allows for comprehension in the more difficult listening condition. We predict that for a more difficult task such as the speeded sentence verification task employed here, participants will not comprehend the synthetic speech as well as they comprehend natural speech. Furthermore, it is predicted that an increased sentence compression rate will have a more negative effect on synthetic speech perception than on natural speech perception.

The overall findings in previous literature can be summarized as follows. a) Naturally spoken words and sentences have typically been shown to be more intelligible and comprehensible than synthetic speech; however, depending on the sophistication of the synthesizing device there may be no noticeable difference in speech quality. b) Time-compressed speech is preferred over naturally fast speech, and c) fast speech rates are more difficult to comprehend than normal, conversational speech rates.

While past studies have compared fast natural speech with time-compressed natural speech, there is a research gap with respect to the comparison of time-compressed natural speech with time-compressed synthetic speech. The aim of this study is to address this gap in the literature. Specifically, by manipulating the variable of speech rate, we determine whether synthetic speech is comprehended as well as natural human speech. Secondly, we establish whether or not comprehension deteriorates equally for both speech signals as the speech rate increases. Based on these results, we assess the quality of current (2010) text-to-speech generators that come pre-installed with Windows and Macintosh computers. Since much of the research on synthesized speech took place over 20 years ago, we predict that the quality of synthesized speech will have improved. If the sophistication of text-to-speech generators has significantly improved, we predict that the participants who are presented with synthetic time-compressed speech will not perform significantly better[1] or worse than those who are presented with natural time-compressed speech. Conversely, if the quality of speech synthesizers has not improved over the last two decades, we predict that the participants who are presented with synthetic time-compressed speech will perform worse than those who are presented with natural time-compressed speech.

[1] “Better” is quantified as a faster reaction time and higher percent accuracy for the speeded sentence verification task.

2 Materials and methods

2.1 Participants

Twenty-five participants (12 male, 13 female) took part in the study. All participants were native Canadian English speakers between the ages of 18 and 30. They reported having limited linguistics training, no major previous exposure to time-compressed or synthesized speech and no hearing loss, although no audiometric test was given. All participants gave their written informed consent to participate in the study, and were not paid or compensated for their time.

2.2 Speech stimuli

This experiment included two sets of auditory stimuli: one synthetic speech set and one natural speech set. Each stimuli set contained recordings of the same 96 true-or-false sentences adapted from Baddeley, Emslie, and Nimmo-Smith's (1992) Speech and Capacity of Language Processing Test, or SCOLP, which was used by both Adank and Janse (2009) and Adank and Devlin (2010). The sentences were slightly altered from their original format: new subjects and predicates were substituted for the original content words. The substituted lexical items were all common, high frequency English words. Only high frequency English words were used in order to avoid a possible confound stemming from lexical confusion. Although the sentential content was altered, the general format of the SCOLP sentences was preserved because SCOLP sentences have been widely tested and have proven to be a reliable measure of language comprehension (Adank & Janse, 2009). A complete list of the sentence stimuli used in the present study is given in the Appendix.

The statements made in the sentences were all unambiguously true or false (e.g., "An ant is a small insect." versus "Elephants are small insects.") in order to ensure that each statement was verifiable. Each true sentence had a false sentence counterpart, as demonstrated in the above example, thus 48 pairs of sentences were used in the experiment. All of the sentences were 7 or 8 syllables long, in order to avoid a possible confound of variable sentence length. A number of past studies have used syllables as the unit of measurement when controlling for sentence length or for the speed of sentence presentation (Adank, & Devlin, 2010; Adank & Janse, 2009; Dupoux & Green, 1997; Janse, 2004).

Of the 96 sentences, 16 were used for pre-test training. The remaining 80 sentences were divided into 5 experimental blocks, with each block containing 16 trial sentences, as is shown below in Table 1. These 16 trials were semi-randomized within their respective blocks. There were an equal number of true and false statements within each block, and sentence pairs were distributed across blocks so pairs did not occur together. The sentences were linearly time-compressed to five different percentages of their original duration using Praat's Lengthen **(**Add-Overlap) function under "Synthesize > Convert" (Boersma and Weenik, 2009). The compression rates used were: 42%, 40%, 38%, 36%, and 34% of the original sentence durations.

Table 1. Experimental Design. Each block contained 16 sentences, followed by 3000 m.s. to respond. Once the answer was recorded, there was a 100 m.s. silence before the next sentence began. Compression rates (%) ranged from 44% to 34%.

Practice (44%)	Block 1 (42%)	Block 2 (40%)	Block 3 (38%)	Block 4 (36%)	Block 5 (34%)

These compression rates were selected on the basis of past research and a participant pre-test. It was decided that each block should include 16 trial sentences because past research has demonstrated that comprehension of a rapid or unusual signal improves over time, and that normalization typically stabilizes after 14–18 sentences worth of experience with a given signal (Adank & Devlin, 2010). The researchers wanted to allow participants a sufficient number of trials at each compression rate so that participants could reach near optimal comprehension performance.

Because the aim of the present study is to compare the comprehension of time-compressed synthesized speech with the comprehension of time-compressed normal speech, two versions of the same 96 sentences were created, one normal speech version and one synthesized speech version. A monolingual female speaker of English from Summerland, British Columbia, Canada recorded the natural speech stimuli. Her recordings were made in a sound-attenuated booth using an M-Audio Luna microphone from the large diaphragm condenser family. The synthesized versions of the experimental stimuli were generated using the text-to-speech "Anna" (Microsoft Inc.) voice that comes included with Windows 7- and Windows Vista-equipped computers. These synthesized sentences were externally recorded with an M-Audio Microtrack solid state recorder. Both the natural and synthetic sentences were clipped to have zero seconds of silence before and after the utterance and saved into 192 separate files. The files recorded by "Anna" were time-compressed or enhanced to be equal length to their natural spoken counterpart.

2.3 Procedure

The study was conducted in the University of Victoria Phonetics Laboratory. All participants received oral instructions read from a script before the experiment began. Participants were randomly assigned to either the synthetic or the natural speech condition. Group A heard natural stimuli, while group B heard synthetic stimuli. The tasks for each group were the same, as were the sentences in each set. This is in accordance with the atypical block-design taken directly from Adank and Devlin (2010). The experiment was run on the software program E-Prime (Schneider et al., 2002a, 2002b). Participants heard the sentence stimuli through headphones at a comfortable sound level, which they determined.

As previously mentioned, the current study will replicate the atypical block design employed by Adank and Devlin (2010). An atypical block design requires that each participant be tested with only one of the two possible signal types. This design is necessary because it has been shown that continually alternating signal type limits behavioral adaptation, thus preventing participants from reaching their optimal performance level (Adank & Devlin, 2010). Because the goal of this experiment is to test the upper limits of synthetic and natural fast speech comprehension, any inhibition of adaptation would be detrimental to the study. Thus, participants were tested on the normal or the synthesized speech signal only.

The participants were first presented with 16 familiarization sentences. The task of the participant was to decide on the validity of each sentence statement presented, and indicate their true-or-false response as quickly as they could with a keyboard button press. Reaction times longer than 3000 m.s. were coded as ‘no response’ and E-Prime automatically presented the next sentence token in the sequence. Both accuracy and reaction time measurements were recorded for each sentence trial. Both measurements were recorded in order to capture in greater detail the cognitive processing costs required for comprehending synthesized and normal speech signals at different compression rates. Reaction time measurements were taken from the end of the audio file following similar previous research procedure (Adank & Janse, 2009; Adank & Delvin, 2010). Good signal comprehension is defined as a high level of response accuracy and short reaction times because these behaviors indicate that the participant was able to easily comprehend and respond to the stimulus presented (Adank & Delvin, 2010).

2.4 Data analysis

Both response accuracy and response time measurements were used as the dependant measures in this study. A total of 25 participants × 96 trial sentences × 2 measurements per trial = 4800 data tokens to analyze (2400 accuracy tokens, and 2400 response time tokens). Accuracy and response time averages were compared between the two speech type conditions and were analyzed across the five compression rate blocks.

3 Results

Table 2 shows the average reaction times and accuracy percentages for the synthetic and natural speech conditions. Overall, participants had shorter reaction times in the normal speech condition versus the synthetic speech condition. The normal speech condition participants also had a higher level of accuracy in their sentence verification responses. Taken together, these two findings suggest that there is a main effect of speech type on comprehension; normal human speech is easier to comprehend than synthetic speech.

Table 2. Average reaction times and percent accuracy across all five blocks.

	Reaction time (m.s.)		Accuracy (%)	
	Normal	Synthetic	Normal	Synthetic
Average:	1015.01	1370.08	85.4	62.5

Figure 1 plots participants' average response accuracy (*y* axis) in making a true or false decision as a function of the signal's compression rate/speed (*x* axis). Average response accuracy for participants in the normal speech condition (black bar graphs), are plotted against the average response accuracy of participants in the synthetic speech condition (grey bar graphs). Figure 1 shows that participants responded more accurately in the normal speech condition than in the synthetic speech condition, for all of the five different compression rates. In the normal speech condition, the lowest response accuracy average was 83.9% and occurred in Block 3 at a 38% compression rate. In the synthetic speech condition, the lowest response accuracy average was 62.5% and occurred in Block 5, at a 34% compression rate. In the normal speech condition, average accuracy rose and fell in random fashion across blocks; there did not appear to be a main effect of compression rate on response accuracy. In the synthetic speech condition average accuracy rose and fell as it did in the natural speech condition, however, there was a general trend that participants in the synthetic speech group became less accurate in their responses as the compression rate was increased.

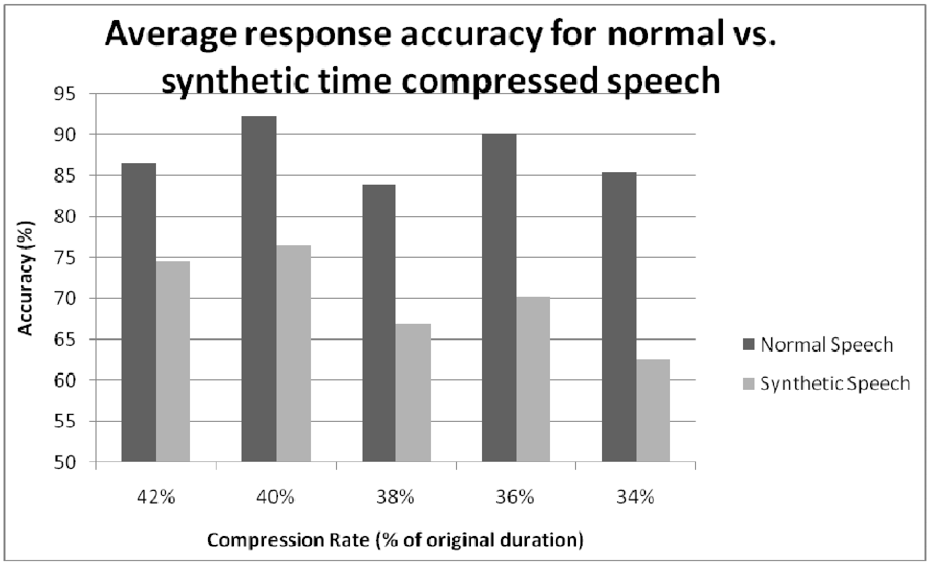

Figure 1. Average response accuracy plotted as a function of speech type and signal speed (compression rate).

Figure 2 plots participants' reaction times (*y* axis) in making a true or false decision as a function of the signal's compression rate/speed (*x* axis). Reaction times for participants in the normal speech condition (black bar graphs) are plotted against the reaction times of participants in the synthetic speech condition (grey bar graphs). Figure 2 suggests that participants in the synthetic speech condition required a longer amount of time to make a sentence verification decision than did the participants of the normal speech condition. When analyzing reaction time performance across compression blocks, we see that in the synthetic speech condition participants' reaction times became steadily slower as the compression rate of the signal increased. In a general way, this effect was also seen in the normal speech condition as well. Figure 2 suggests that there is a main effect of compression rate on decision response time, and that the normal speech signal is easier to perceive than the synthetic speech.

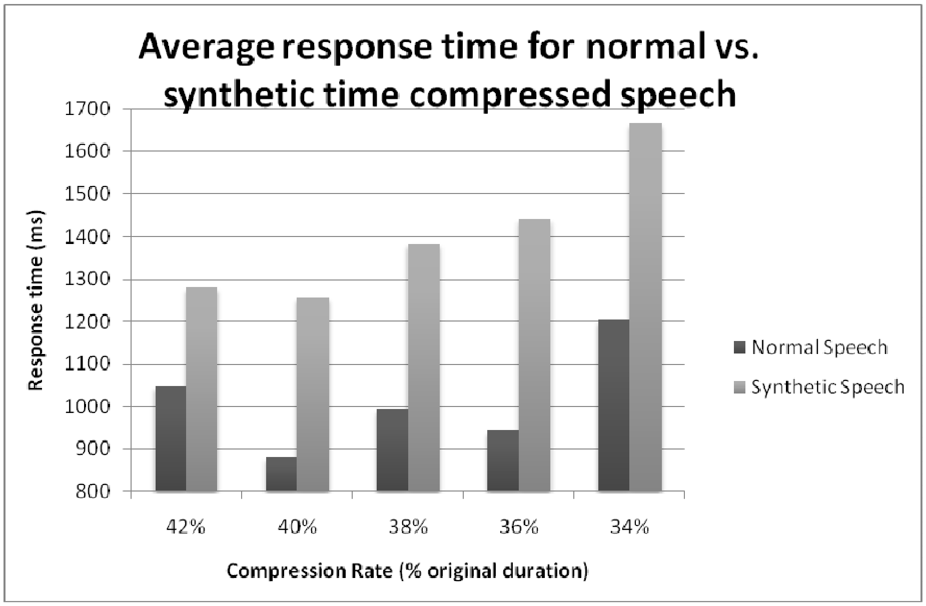

Figure 2. Average reaction time plotted as a function of speech type and signal speed (compression rate).

4 Discussion

4.1 Key findings

The results demonstrate two important points. First, they show that listeners are not able to comprehend synthetic speech as well as they comprehend natural speech. This goes against Koul (2003), who found synthetic speech to be of comparable quality to natural speech, although the majority of the literature supported the view that synthetic speech is more difficult to perceive and requires greater cognitive resources to comprehend (Duffy & Pisoni, 1992; Francis & Nusbaum, 2009; Mirenda & Beukelman, 1987). Listener performance in the present experiment confirms that synthetic speech is more difficult to process than natural speech. Individuals in the synthetic speech condition had longer reaction times and lower response accuracy averages than normal speech participants for all five of the compression rates.

Secondly, the results show that the adverse listening condition of fast speech makes comprehension more difficult for synthetic speech listeners than for natural speech listeners. Both the average accuracy results and the reaction time results support this finding. In both speech conditions, the average accuracy of participant responses rose and fell as the compression rate was increased. In the normal speech condition, participants' average accuracy for Block 1 (42% compression) was 86.5% and their average accuracy for Block 5 (34% compression) was 85.4%. This indicates that increasing the speed of natural speech does not have a large effect on participants' ability to accurately verify sentences. In the synthetic speech condition, however, increasing the speed of the signal did affect comprehension performance. The average response accuracy for the synthetic speech group in Block 1 was 74.5%, and this already low figure dropped to 62.5% accuracy by Block 5, a difference of -10%. The reasons why synthetic speech comprehension may have been so adversely affected by a fast signal presentation rate will be discussed in detail below.

The reaction time results also suggest that the comprehension of synthetic speech is more affected by an increased signal rate than normal speech. In the normal speech condition, the average reaction time of participants rose and fell across blocks, although the general trend was that reaction times became longer as the presentation rate increased. For the normal speech condition, Block 2 (40% compression) had the shortest average reaction time of 881.00 m.s. Block 5 (34% compression) had the longest average reaction time of 1204.89 m.s. The difference between the fastest and slowest reaction time averages for participants in the natural speech group was +323.89 m.s. The shortest average reaction time in the synthetic speech condition occurred in Block 2 (40% compression) and was 1256.59 m.s.; the longest reaction time average occurred in Block 5 and was 1666.40 m.s., a total difference of +409.81m.s. The fact that there is a larger reaction time difference for the synthetic speech group than for the natural speech group suggests that participants in the synthetic speech condition were more adversely affected by the increase in speech rate.

In sum, our results show that listeners are not able to comprehend synthetic speech as well as they are able to comprehend natural speech, and that an increase in speech rate has a greater adverse affect on synthetic speech perception than on natural speech perception. These findings are in line with our original hypotheses. Despite the technological advances that have greatly improved the quality of synthetic speech in recent years, there are a variety of possible reasons why people are still unable to comprehend synthetic speech as well as they comprehend natural human speech. First, let us consider the Windows 07's "Anna" voice that was used in the present study. The Microsoft "Anna" voice was created with formant synthesis technology. In formant synthesis, the different acoustic parameters of speech such as fundamental frequency, voicing, and signal amplitude, et cetera, are produced by algorithmic rules, which create the artificial speech waveform. For this type of speech synthesis, it is common that only one or two acoustic cues will be specified to distinguish a given phoneme, and often the same acoustic cues are used for more than one phoneme. Researchers Francis and Nusbaum (2009) identify this impoverished and misleading cue structure as being the primary reason why synthetic speech perception can be so difficult. In natural speech, there are multiple acoustic cues that interact to create the percept of a specific phoneme. In synthetic speech, on the other hand, perceptual ambiguity may be increased because (1) fewer discrete cues have been encoded, so the relationships between the synthesized acoustic cues may be uninformative and misleading in comparison with the cue structure of natural speech, and (2) the same patterns of acoustic cues appear in a greater range of contexts for synthesized speech (Francis & Nusbaum, 2009). This acoustic-phonetic ambiguity, which is found in synthetic speech, is one possible reason why synthetic speech comprehension is difficult. Also, because speech synthesizers are engineered by humans, there is always the possibility that human engineering errors could result in misleading cue structure (Francis & Nusbaum, 2009). In such circumstances, the listeners would need to learn to inhibit their perceptual intuitions for the poorly engineered contexts in question.

Another possible reason why the synthetic speech stimuli may have been more difficult to comprehend is that this study tested for listeners’ comprehension of synthetic speech rather than just the intelligibility of the signal itself. Comprehension requires a higher level of cognitive processing than does simple perception because comprehension involves perception, acoustic-phonetic mapping, and lexical access (Koul, 2003). In fact, even for high quality synthetic speech, a substantial portion of cognitive resources are allocated to deciphering the acoustic-phonetic structure of the signal, leaving fewer resources available for higher level semantic processing (Duffy & Pisoni, 1992). Because a speeded sentence verification task is a relatively complex task, it is possible that participants’ cognitive resources were focused on low level perception and thus unable to efficiently construct a mental representation of the message. If this were the case, such findings would have important implications for speech-generating-devices and for the individuals who use them.

4.2 Limitations

A limitation of this study is that true-or-false sentence pairs were used for the experiment stimuli. The 96 sentence pairs used were all altered SCOLP sentences. SCOLP format sentences were chosen because the SCOLP test has been proven to be a reliable measure of language comprehension (Adank & Janse, 2009), and because similar studies involving linearly time-compressed speech had used these sentences in the past (Adank & Janse, 2009; Adank & Devlin, 2010). Unfortunately, many of our participants reported that after they had gained some experience with the speeded sentence verification task, they realized that the sentence stimuli were arranged into pairs, (e.g. “Governors work in politics.” vs. “Strawberries work in politics.”) and that one member in the pair would always be true and the other would always be false. This awareness enabled some participants to respond faster for the second sentence in a pair – they exhibited a repetition priming effect. The decrease in reaction time and increase in accuracy, which accompanied their repetition priming effect, meant that some participants performed better as they became increasingly familiar with the words used in the sentences, and with the sentences themselves. This effect counter-acted the decrease in comprehension that was predicted to occur as the speech signals became increasingly fast. If many words are initially recognized, then it is relatively easy to engage in a guessing strategy that reconstructs the initially unintelligible words (Dupoux & Green, 1997). Thus, for the second sentence in a pair it is possible that guessing strategies had a larger effect on response accuracy and reaction time than did actual signal comprehension. Future trials of this experiment could address this deficiency by continuing to use obviously true or false sentences for verification, but ensuring that each sentence occurs in isolation with no semantically related pair item.

Another possible limitation of this study is that the compression rates used were not as widely distributed as they perhaps should have been. Recall the five different linear time compression rates employed in this study: 42%, 40%, 38%, 36% and 34% of the original sentence durations. Dupoux and Green (1997) found that sentences compressed to 38% of their original duration were difficult for participants to understand, while Adank and Devlin (2010) found that listener's required 10–20 sentences to adapt to material that had been compressed to 35% of its original duration. In light of the contradictory past research, a pre-test was administered to 3 participants in order to determine a suitable range of compression rates. Participants in the pre-test heard eight sentences at each of the seven possible compression rates: 44%, 42%, 40%, 38%, 36%, 34% and 32%. Participants were seated in a quiet room and the sentences were played over a loudspeaker for all to hear. The pre-test participants exhibited excellent comprehension at the 44% compression rate and substantial difficulties in sentence comprehension starting at the 36% compression rate. On the basis of the pre-test participants' performance, it was decided that a 44% compression rate would be used for the training stimuli and that a 38% compression rate should be the median experimental compression value. We predicted that participant comprehension in Blocks 1 and 2 (42% and 40% compression) should be quite good as these two rates are slower than the median 38% value. We similarly predicted that comprehension in Blocks 4 and 5 (36% and 34% compression) should be quite poor as these two rates were faster than the selected median value. Surprisingly, the experimental participants in both speech conditions exhibited high comprehension throughout the experiment. Even in Block 5, the fastest compression rate presented, participants in the synthetic speech group still performed at above chance level (62.5%) for sentence verification accuracy. Future trials of this experiment could address this deficiency by using a broader range of compression rates so as to better delineate the relationship between compression rate and comprehension. Furthermore, if a pre-test is administered, stimuli should be delivered in the same way (e.g. loudspeaker, headphones) as it will be delivered in the experiment.

5 Conclusion

In conclusion, our results have shown that despite recent advances in formant synthesis technology, listeners are still unable to comprehend synthetic speech as well as they comprehend natural human speech. Additionally, the comprehension of synthetic speech is more affected by adverse-listening conditions such as increased speech rate than is natural speech. Because text-to-speech generators play an important helpful role in the lives of visually and communicatively impaired individuals, and are widely used in the fields of language translation, business, and entertainment, these results are highly relevant. They indicate that further work is needed to improve the quality of synthetic speech for the sake of all individuals who use such signals. Our results thus add to the sizable body of research on synthetic speech perception. The researchers suggest that a similar study, which uses time-compressed speech to compare the quality of many different text-to-speech generators, would be beneficial to this field.

References

Adank, P. & Devlin, J.T. (2010). On-line plasticity in spoken sentence comprehension: Adapting to time-compressed speech. *NeuroImage* 49(1), 1124–1132.

Adank, P. & Janse, E. (2009). Perceptual learning of time-compressed and natural fast speech. *Journal of the Acoustical Society of America* 126(5), 2649–2659.

Baddeley, A.D., Emslie, H. & Nimmo-Smith, I. (1992). *The Speed and Capacity of Language Processing (SCOLP) test*. Bury St Edmunds: Thames Valley Test Company.

Boersma, P. & Weenink, D. (2009). Praat: doing phonetics by computer (Versions 4.5). [Computerprogram] http://www.praat.org/.

Duffy, S.A. & Pisoni, D.B. (1992). Comprehension of synthetic speech produced by rule: A review and theoretical interpretation. *Language and Speech* 35(4), 351–389.

Dupoux, E. & Green, K. (1997). Perceptual adjustment to highly compressed speech: Effects of talker and rate changes. *Journal of Experimental Psychology: Human Perception and Performance* 23(3), 914–927.

Francis, A.L. & Nusbaum, H.C. (2009). Effects of intelligibility on working memory demand for speech perception. *Attention, Perception, and Psychophysics* 71(6), 1360–1374.

Golomb, J.D., Peelle, J.E. & Wingfield, A. (2007). Effects of stimulus variability and adult aging on adaptation to time-compressed speech. *Journal of the Acoustical Society of America* 121(3), 1701–1708.

Janse, E. (2004). Word perception in fast speech: Artificially time-compressed vs. naturally produced fast speech. *Speech Communication* 42(2), 155–173.

Koul, R. (2003). Synthetic speech perception in individuals with and without disabilities. *Augmentative and Alternative Communication* 19(1), 49–58.

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(2), 112–122.

Koul, R. & Hanners, J. (1997). Word identification and sentence verification of two synthetic speech systems by individuals with intellectual disabilities. *Augmentative and Alternative Communication* 13(2), 99–107.

Koul, R. & Hester, K. (2006). Effects of repeated listening experiences on the recognition of synthetic speech by individuals with severe intellectual disabilities. *Journal of Speech, Language, and Hearing Research* 49, 47–57.

Mirenda, P. & Beukelman, D.R. (1987). A comparison of speech synthesis intelligibility with listeners from three age groups. *Augmentative and Alternative Communicaation* 3(3), 120–128.

Pallier, C. & Sebastian-Galles, N., Dupoux, E, Christophe, A. & Mehler, J. (1998). Perceptual adjustment to time-compressed speech: A cross-linguistic study. *Memory & Cognition* 26(4), 844–851.

Schneider W, Eschman A & Zuccolotto A. (2002). *E-Prime reference guide*. Pittsburgh: Psychology Software Tools Inc.

Appendix

The number of syllables in each sentence is listed to the right of the sentence.

	Set 1			Set 2	
1.	Beavers build dams in the river.	8	1.	Governors work in politics.	8
2.	2. A tomato grows on a plant.	8	2.	Monks live in a monastery.	8
3.	3. Telephones can be bought in stores.	8	3.	Shovels are used in the garden.	8
4.	4. Motorcycles drive on the road.	8	4.	Sirloin steaks are sold by butchers.	8
5.	Fish breathe oxygen through gills.	7	5.	A leopard has a fur coat.	7
6.	Donkeys carry heavy loads.	7	6.	Butterflies have antennae.	7
7.	Carrots grow in a garden.	7	7.	A butcher works in a shop.	7
8.	An architect has a job.	7	8.	Wool is made from a sheep's coat.	7
9.	A camel is a kind of bird.	8	9.	Eagles build dams in the river.	8
10.	Dishwasher fluid walks the streets.	8	10.	A rainbow trout grows on a plant.	8
11.	Fathers are stored in the toolbox.	8	11.	Oxygen can be bought in stores.	8
12.	Biking is slower than walking.	8	12.	Fresh lemonade drives on the road.	8
13.	Buddhism is a pencil box.	8	13.	Pigs breathe oxygen through gills.	7
14.	Backpacks are always women.	7	14.	Babies carry heavy loads.	7
15.	Elephants are small insects.	7	15.	Beavers grow in a garden.	7
16.	April is a summer month.	7	16.	A vegetable has a job.	7

	Set 3			Set 4	
1.	A cake is baked in an oven.	8	1.	A tank is a weapon of war.	8
2.	Elephants are living beings.	8	2.	A minute is sixty seconds.	8
3.	Tables and chairs are furniture.	8	3.	Exercise is good for your health.	8
4.	Wooden chairs are for sitting on.	8	4.	A trout is a species of fish.	8
5.	Geese can fly long distances.	7	5.	A melon is a type of fruit.	8
6.	Bees fly around looking for food.	7	6.	Spoons are used for eating soup.	7
7.	A captain commands the ship.	7	7.	A shed is used for storage.	7
8.	Knives are used in the kitchen.	7	8.	Wine bottles are made of glass.	7
9.	A bike is a weapon of war.	8	9.	Strawberries work in politics.	8
10.	An hour is forty minutes.	8	10.	Donkeys live in a monastery.	8
11.	Smoking is very good for your health.	8	11.	A cake is used in the garden.	8
12.	An ant is a species of fish.	8	12.	Architects are sold by butchers.	8
13.	A cabbage is a type of fruit.	8	13.	A goldfish has a fur coat.	7
14.	Forks are used for eating soup.	7	14.	Bathroom sinks have antennae.	7
15.	Nurses are used for storage.	7	15.	A lion works in a shop.	7
16.	Policemen are made of glass.	7	16.	Ink is made from a sheep's coat.	7

	Set 5	
1.	A pelican is a bird species.	8
2.	Police officers walk the streets.	8
3.	Hammers are stored in the toolbox.	8
4.	Walking is slower than biking.	8
5.	Buddhism is a religion.	8
6.	Mothers are always women.	7
7.	An ant is a small insect.	7
8.	August is a summer month.	7
9.	Dentists are baked in the oven.	8
10.	Cabinets are living beings.	8
11.	The plastic doll is furniture.	8
12.	Computers are for sitting on.	8
13.	Grapes can fly long distances.	7
14.	Flies walk around looking for food.	7
15.	A leopard commands the ship.	7
16.	Snakes are used in the kitchen.	7