The Effect of Task Conditions on the Comprehensibility of Synthetic Speech

Jennifer Lai, David Wood
IBM Corporation/T.J. Watson Research Center
30 Saw Mill River Road
Hawthorne, New York 10532
lai, dawood@watson.ibm.com

Michael Considine
Rice University
6300 S. Main Street
Houston, Texas 77005
consid24@rice.edu

ABSTRACT
A study was conducted with 78 subjects to evaluate the comprehensibility of synthetic speech for various tasks ranging from short, simple e-mail messages to longer news articles on mostly obscure topics. Comprehension accuracy for each subject was measured for synthetic speech and for recorded human speech. Half the subjects were allowed to take notes while listening, the other half were not. Findings show that there was no significant difference in comprehension of synthetic speech among the five different text-to-speech engines used. Those subjects that did not take notes performed significantly worse for all synthetic voice tasks when compared to recorded speech tasks. Performance for synthetic speech in the non note-taking condition degraded as the task got longer and more complex. When taking notes, subjects also did significantly worse within the synthetic voice condition averaged across all six tasks. However, average performance scores for the last three tasks in this condition show comparable results for human and synthetic speech, reflective of a training effect.

KEYWORDS
Text-to-Speech, Synthetic Speech, User Study, Comprehension

INTRODUCTION
For many years now, people have been predicting that the use of computer generated speech would become widespread; showing up in applications ranging from talking appliances and vending machines, to fully conversational data entry and retrieval systems. For the most part however it seems that the forecasted wide-ranging use of synthetic speech has not materialized, and that we encounter it primarily in situations where the use of recorded human speech is not practical, as in access to large and frequently changing databases of information such as stock quotes or mutual fund prices. This may be because people report that synthetic speech sounds unnatural and is unpleasant to listen to [1, 7, 8]. More recently though, synthetic speech has been applied to the retrieval of e-mail messages and the reading of news articles in products such as Portico by General Magic [10] and Webley by Webley Systems Incorporated [11].

Synthetic speech, also known as text-to-speech (TTS), is speech produced by a computer. It is often referred to as "rule-based" speech because the computer uses a series of rules to convert text to the sounds that are generated. The text goes through several stages of transformation prior to the actual synthesis itself. The type of synthethis varies from one TTS engine to another and can be formant-based, articulation-based, or concatenative [3]. Such systems are capable of producing an unlimited number of messages without storage constraints, and are absolutely necessary when reading dynamically created text such as e-mail messages and late breaking news stories.

As part of a group of researchers in Pervasive Computing Solutions, facing the prospect of incorporating synthetic speech into several of our prototypes, my colleagues and I have a keen interest in measuring how well synthesized speech can be understood. One of the areas in which we are focusing our research involves the use of a virtual assistant, which would allow users to call from any telephone and stay in touch with their office by interacting with an "assistant" which conveys information about messages and select news articles. Although we believe that digitally recorded human speech would be subjectively preferred over synthetic speech in such an application, the dynamic nature of the messages and file size considerations make this solution impractical if not impossible.

Our goal was not to produce a rank-ordering of the five synthetic speech engines measured, but to understand if there were optimal conditions for the comprehension of synthesized speech, and to what degree comprehension would degrade when these conditions varied. The conditions we looked at were the nature of the task and the effect of note-taking while listening. The tasks consisted of

short informal e-mail messages, longer messages, and CNN news stories. We used comprehension levels of the passages when read by a professional voice talent as a reference condition for each subject.

PRIOR WORK

Most of the prior evaluations of text-to-speech systems have concentrated on either segmental intelligibility (at the word or sentence level) or listener preference (subjective evaluation of the acceptability of synthetic speech). The two tests most often used for word level intelligibility are the Modified Rhyme Test (MRT) and the Diagnostic Rhyme Test (DRT) [1]. Both are closed response tests where the user hears a word and is asked to select the word from a list of possible choices. Sentence level intelligibility has been measured by using either the Harvard Psychoacoustic sentences or the Haskins Syntactic sentences [8]. The Harvard sentences are a fixed set of semantically and syntactically normal sentences such as "*Add salt before you fry the egg*". The Haskins sentences are also a fixed set of 100 sentences, but in this case they are semantically unpredictable as in "*The old farm cost the blood*". Although findings indicate synthetic speech is less intelligible than natural speech and that it takes longer to process [5, 6], a wide variance in the performance of synthesizers has been reported.

Even though segmental intelligibility is a critical component in the overall acceptability of a speech synthesizer, Ralston et. al [7] point out the significant difference between the task of recognizing words and sentences, and the task of understanding the intended meaning of the message or spoken passage. When listening to longer passages of synthetic speech, issues of context, task, prosody, fluency (timing) and intonation come into play. These issues may have either a negative impact on comprehension or a positive one, based on the task and synthesizer.

Very few studies have assessed the comprehension of synthetically produced passages of meaningful continuous speech. One study often referred to in the literature, by Pisoni and Hunnicutt in 1980 [4], measured comprehension based on 15 narrative passages of text obtained from a standardized adult reading test. Comprehension performance was measured based on the ability to correctly answer multiple choice questions when listening to synthetic speech, natural speech and reading the text. They found that subjects did significantly better if reading the passage rather than hearing it with either natural or synthetic speech. This effect may have been skewed by the fact that the passages were designed to measure reading comprehension, not auditory comprehension. The study also showed that subjects did significantly better on the second half of the test when listening to synthetic speech, a gain that was not seen in the other media. Overall comprehension levels for synthetic speech were comparable to natural speech for the MITalk synthetic speech system (the only TTS system measured in this study).

We were motivated to conduct a study on the comprehensibility of synthetic speech for several reasons. The first is that most of the findings reported in the literature evaluate the perception of synthetic speech based on TTS engines from 1979 to 1986. Given the progress in technology over the past 13 years we believed it would be valuable to reassess comprehension levels. The second reason is that we agree with Ralston et. al [7] who indicate after a comprehensive review of the research over the past 15 years, that evidence in the matter of the comprehension of synthetic speech is not totally compelling. While a body of findings indicate reliable differences in levels of comprehensibility between synthetic and natural speech, results vary considerably across different studies. Our third motivation was a desire to understand if certain tasks were better suited than others for listening to with synthetic speech on the phone. We structured our study based on findings in Pisoni et al. [5] that there are five major factors that influence listener's performance in laboratory situations 1) quality of speech signal 2) size and complexity of message set 3) short-term memory capacity of listener 4) complexity of the listening task or other concurrent tasks 5) the listener's previous experience with the system.

SYSTEMS

Five commercially available text-to-speech systems were used in the study. While it was tempting to use more "cutting edge" prototype systems available either from Universities or Research organizations, the decision was made to draw the line at commercial products. It was not possible given constraints on time and resources, to include every commercial TTS engine available in the market today. Ultimately we decided to use: DECtalk for Windows 95 v 4.4, AcuVoice AV1700 text reader, IBM Via Voice Outloud (from VV '98), L&H TTS engine version 6.03, and Lucent Release 2. The first two products were purchased, the IBM product was acquired internally, and the latter two were sent to us as evaluator products, free of charge.

METHOD

Experimental Design

Each subject was randomly assigned to one of the five text-to-speech products. Half the subjects heard synthesized speech first and half heard recorded human speech first. The order of passage presentation was counterbalanced using a Latin Square design. Dependent measures included: (1) response accuracy, and (2) time on task (the time from when a subject turned over the questionnaire to view the questions, until completion of responses). Digitized natural speech was used a reference condition for every subject.

Subjects

All 78 subjects were IBM employees and native English speakers recruited from within IBM Research or the neighboring building which houses the Internet division. Subjects were screened to ensure that they did not report any prior history of hearing problems and that they were not working in the field of synthetic speech or otherwise overly familiarized with text-to-speech. Half of the subjects were female and half were male. All subjects volunteered for the study and received a $20 gift certificate for their participation in the 45 minute session. Consent was obtained prior to videotaping the session.

Procedure

To familiarize subjects with the task at hand, each subject was given a practice task prior to experiment participation. The training passage was recorded using a human voice, so that no expertise could be acquired with synthetic speech prior to the actual test itself. A human voice other than the one used in the experiment was used for the training passage so that there too, there would be no adaptation or familiarization. After hearing the training passage and answering the practice questions the subjects were given the correct answers to self-check their work. They were asked if they had any questions about the procedure and if the volume was appropriate for them.

The subject was then instructed to begin the test. Listening to the test passages required following the same process the subject had used for the training passage, namely dialing a four digit telephone number (which was posted next to the phone), listening to the text, hanging up the phone, and turning over the sheet of questions for the text. Thus, the subject had no knowledge of the questions prior to hearing the text, and was not allowed to hear any portion of the passage again. The subject was given as much time as needed to answer as many questions as possible. The amount of time spent answering the questions was recorded for each passage. Once the subject was done with a passage, the process started over again until all six test tasks were completed.

The test consisted of two sets of data. Each set was made up of a short (S), a medium (M) and a long (L) passage. The passages in each set were similar but different. Each subject heard one set "read" by a synthetic engine, and a set with a recorded human voice. The order of voice presentation was alternated such that subject N heard set 1 with human voice and set 2 with synthetic voice and subject N+1 heard set 1 with synthetic and set 2 with human. As mentioned earlier, the order of passage presentation within each set was counterbalanced using a Latin Square. The first 36 subjects always heard set 1 first (regardless of the voice type) and were allowed to take notes. The next 36 subjects always heard data set 2 first and were not allowed to take notes.

The comprehension questions consisted of multiple choice questions of both general comprehension, and specific recognition.

Sample general comprehension question:

2) What is the main topic of this story?

a.) Indian Foreign Minister Singh's request to endorse the inviolability of the cease fire line;
b.) Pakistan's attempts to disguise their support of Islamic militants;
c.) The breaching of the 1972 cease-fire line;
d.) The tenuous peace between Pakistan and India

Sample specific recognition question:

1) What is the room number for the meeting?

a. H1-D20
b. H1-D30
c. H3-D20
d. H2-B30

One quick word on our choice of an offline/successive measure rather than a simultaneous measure such as word monitoring. A successive measure is one that is made after presentation of the material, while a successive test measures comprehension as it is taking place [7]. We opted for multiple choice questions which are one of the oldest and most commonly used recognition tests for comprehension. While it does not allow for measuring of processing load during comprehension, it is a good test for determining what has been extracted and retained from the text passage. This choice corresponds to our interest in understanding how well pertinent information presented in an application can be remembered.

Task

Six passages of text were selected for the study. Our goal was twofold in the selection of passages. The first was to match the text in the test to the type of text we expect will be used in the virtual assistant (VA) application. These are primarily e-mail messages of varying length and news stories. The second goal was to select from standardized tests created to measure listening comprehension, not reading comprehension. To this end, we had three sources for the three types of tasks. The short passages were approximately 100 words long and consisted of informal and fictitious e-mail messages. The medium passages (about 300 words) were derived from the Test of English as a Foreign Language (TOEFL) and the long passages (about 500 words) from CNN news stories.

The TOEFL [9] is an English proficiency standardized exam, which has sections on listening comprehension, written expression and reading comprehension. The listening comprehension section is divided into three parts: the first consisting of short conversations, the second of

longer conversations and the third of short talks. We selected our passages from the short talks. We took talks from sample exams, that could conceivably be received as messages or obtained as information from the Web, such as weather reports, or biographical information on guest speakers.

The CNN stories that we selected were chosen as ones which we believed would be mostly unfamiliar material for many of our subjects. These were stories which only "news hounds" or people with a particular interest in the subject matter would be closely familiar with. The questions were designed so that even a person generally familiar with the topic, would be unable to answer them correctly without having heard the particular article. For example:

1. Which is NOT mentioned in this article as a reaction to the verdict?

a.) security was reinforced at airports, prisons and embassies;

b.) certain activists were placed under surveillance;

c.) violent protest broke out in several European cities;

d.) Turks sang the national anthem;

As a check for this, we had 10 subjects, independent of those used in the comprehension study answer all six sets of questions without ever hearing the text. In spite of all these precautions however, we realized that some subjects could have more familiarity with the news articles than others. We felt that this would also be true of users listening to newstories in the virtual assistant, and since a goal of ours was to recreate conditions of application usage, we decided to include real news articles instead of fictitious ones.

We expected that listeners would be able to apply a top-down real-world knowledge to deciphering in context, words and names which might have otherwise been unintelligible. For example the word "mudanya" would be difficult to perceive in a word level intelligibility test, but is easy to comprehend as part of the sentence "Families of soldiers killed in fighting against the PKK gathered at Mudanya, the port closest to the prison island where the trial is being held."

Listening Environment

The experiment was carried out in what could be considered optimum listening circumstances, especially when compared to a potential real-world setting such as being dialed in over a cell phone while driving one's car. The subjects were seated alone in the observation area of a usability lab. There were no interruptions or distractions.

Since the conditions of usage of the VA are potentially quite varied (one could call from a phone anywhere) it was difficult to duplicate the environment for testing purposes. However, certain factors that are constant for the VA were reproduced. The subjects were required to call in using a telephone, and listened to text that was of a similar nature to that of the application. The level of experience, motivation and demographic characteristics of the subjects fell into the same range as the user set we believe would adopt usage of this type of application.

For each engine (except AcuVoice) we selected the "adult male" voice and the "adult female" voice, using them with their default settings for pitch and breathiness. With AcuVoice we only had male voices to select from so no female voice was used. We modified, where necessary, the words per minute (wpm) rate so that all engines "read" the text at 175 wpm.

The passages of text recorded by a professional voice talent were timed and were shown to have a rate of 200 wpm. The voice talent we hired is male. The human voice was recorded at a sampling rate of 44,100 and a 16 bit resolution. For the synthetic speech passages, the text was "read" each time rather than being played as a stored .wav file.

Evaluation Measures

Recognition memory was measured at the end of each set by presenting the subject with a list of 20 sentences. Ten of the sentences were actually spoken during the set, the other 10 sentences had semantically the same meaning but were presented with a change in the wording or grammar.
For example:

a. The rival nuclear powers have fought three wars since 1947, two of them over Kashmir.

b. Since 1947 three wars have erupted between these rival nuclear powers, two of which have been fought over Kashmir.

In addition to measuring comprehension performance and time on task, subjective measures were obtained by calculating the Mean Opinion Score [1], the use of an exit questionnaire and exit interview. Here again our goal was not a rank-ordering of the five engines, but to find out if there was a large variance between the subject's perception of his or her performance and the actual level of performance itself. We were also looking for any gender preference issues that might surface.

At the end of the study, subjects were asked to complete a Mean Opinion Score (MOS) survey and exit survey. They were instructed that the MOS was only pertinent to the passages they had heard in synthetic speech. The MOS survey is comprised of seven evaluative 5-point scales measuring:

1. Global Impression; 2. Listening Effort; 3. Comprehension Problems; 4. Voice Pleasantness; 5.Speech Sound Articulation; 6. Pronunciation; 7. Speaking Rate

The first four can be viewed as measuring the overall quality of the speech while the last three scales relate to specific aspects of the output and require more analytical listening.

FINDINGS

There was no significant difference for comprehension performance of synthetic speech among the five different TTS engines used ($F(4,70) = .106$, $p = .980$).

Nor was there an effect found for time on task. The amount of time required to answer questions in the synthetic set was nearly equal to the time required for the human set (see table 1). The difference in time between the note conditions suggests that subjects looked at their notes to a minor degree while answering questions. We speculate that the lack of difference in timings is due primarily to the off-line nature of the multiple choice questions.

	Notes	No Notes
Hum. Av. Time	199 seconds	174 seconds
Syn. Av. Time	202 seconds	175 seconds

Table 1: Average time on task for the set as a function of voice type and note-taking.

The Training Effect

An important factor with TTS is the rate at which humans adapt their ear to the sound of synthetic speech (or human speech when spoken by a foreigner) and the effect that this adaptation has on levels of comprehension. The amount of training necessary to see a change in the perception of synthetic speech ranges in various studies from a few minutes of exposure [8] to four hours of training [2]. Also when subjects improve performance during the course of a study, it is hard to differentiate what portion is due to a familiarization with the process, and what is due to training on synthetic speech.

Subjects in this study showed an ability to quickly learn both cognitive and perceptual strategies to improve performance. As seen in Table 2, subjects within the note taking condition, who heard the human voice first achieved a comprehension performance of 82% for set 1 and 73% for set 2. Those that heard synthetic first scored 59% for set 1 and 67% for set 2.

	Set 1	Set 2
Hum.	82%	67%
Syn.	59%	73%

Table 2: results by set for the note taking condition (N=36) Results are shown as a percentage of correct answers.

Even though the first group was hearing set 2 in synthetic speech, they scored higher on this set than the group that heard it with natural speech. A post-study analysis showed that there were no significant demographic differences among the four groups of subjects in our study.

This slightly counterintuitive finding could be explained by the fact that subjects try harder when listening to synthetic speech and that the task training which had occurred by the second set had a stronger effect in the synthetic condition. We believe that many subjects trained very rapidly, perhaps within the first sentence or so of synthetic speech. One subject commented "I found I'd miss the first sentence while I trained on the voice, so perhaps some introductory preamble would help." Another said that "by the end of the session I found myself less distracted by the idiosyncrasies of the synthesized voice and better able to focus on the content."

The Effects of Age, Education and Prior TTS Exposure

Even though subjects were screened to eliminate any people who had significant prior exposure or interaction with synthetic speech, we still had 13 (5 for the notes condition and 8 for the non note-taking condition) who arrived for the study and indicated on their data sheet that they used it with "some regularity". Each subject was asked to report his or her level of experience with TTS using the following scale:

1. know what it is but have not heard it
2. have heard it once or twice on the radio or TV
3. have cause to listen to it with some regularity
4. hear it at least once a week
5. work with it

No subjects were included that reported prior exposure at levels 4 or 5.

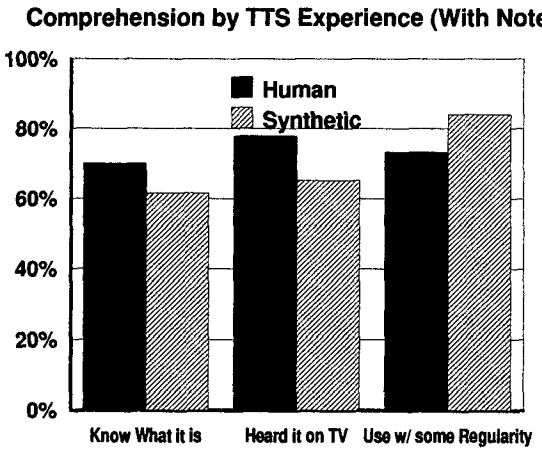

Figure 1: Performance as a function of voice type and experience with TTS for the notes condition.

As Figure 1 shows, there was an interaction between voice type and TTS experience within the notes condition ($F(2, 35) = 5.240$, $p = .01$). Subjects who had cause to listen to TTS "with some regularity" did better for the synthetic voice than for human, as well as doing better than those subjects who had less exposure. The advantage these

subjects had was lost in the non note-taking condition, as shown in Figure 2.

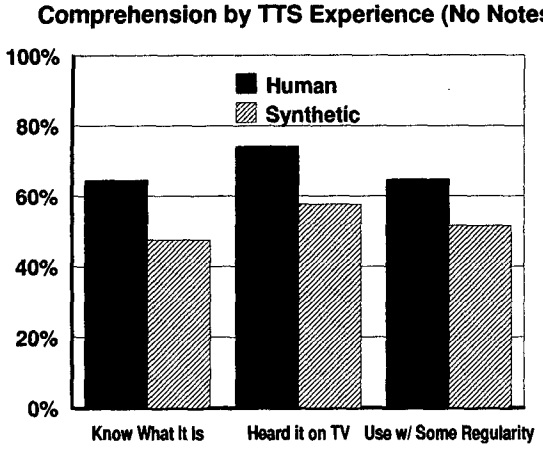

Figure 2: Performance as a function of voice type and experience with TTS for the non note-taking condition.

As expected, subjects who had completed a higher level of education performed better overall ($F(3,71) = 8.200$, $p < .001$). However there were no significant interactions with other variables such as note-taking and voice type. Across both note-taking conditions, there was no effect observed for age of subjects ($F(3,71) = 1.991$, $p = .13$).

The Effect of Note Taking

We observed for the most part, that subjects who were told they could take notes, took a large amount of notes. When answering the questions however, only rarely did we see the subjects referring back to their notes.

While we expected to see significant differences in performance between the three types of tasks, we did not observe this effect within the note taking condition. Certainly the ability to take notes while listening is not reflective of task conditions if listening to messages while driving a car, however we wanted to see what differences would surface that were totally independent of any short-term memory issues.

	S2	M2	L2	S1	M1	L1
Hum.	70%	69%	51%	70%	85%	71%
Syn.	65%	48%	38%	61%	69%	42%

Table 3: performance by task for the non note-taking condition (N=36)

The notation in Table 3 corresponds to the task (S,M,L) and the data set. Thus S2 is the short passage in set 2.
As seen in Table 3, subjects performed worse for all passages with synthetic speech in the non note-taking condition. Performance deteriorated as the task became longer and less familiar.

Comparisons in Figure 3 of comprehension performance for the note-taking and non note-taking the condition show that (perhaps not surprisingly) subjects did better overall when allowed to takes notes ($F(1,73) = 8.026$, $p = .006$). Subjects performed better on the medium ($t(73) = 3.146$, $p=.002$) and long ($t(73) = 2.789$, $p=.007$) synthetic passages in the note-taking condition. What is more interesting is that there was also a significant interaction between note-taking and voice type ($F(1,73) = 4.175$, $p<.05$). This indicates that the effect of note-taking is significantly stronger when listening to synthetic speech rather than human.

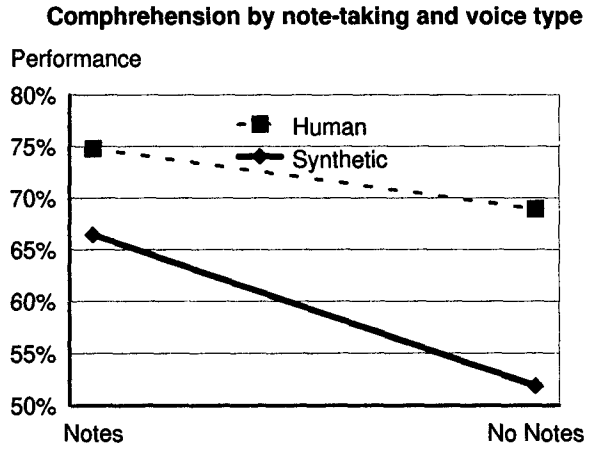

Figure 3: performance by voice type for the notes and non note-taking condition

Results were similar for the recognition memory task. As seen in Table 4, note-taking had no effect on the subjects' ability to recognize phrases with human speech. However, with synthetic speech, recognition memory was improved with notes ($t(63)=2.329$, $p=.023$).

	Notes	No Notes
Hum. Rec. Mem.	70%	70%
Syn. Rec. Mem.	66%	58%

Table 4: Recognition Memory performance as a function of voice type and note-taking

Gender Issues

MOS scores for the four engines that had both male and female voices, showed that female subjects showed a slight preference for the male voice over the female voice (preference seen in 3 out of 4 engines). There was no clear preference for male subjects.

Although there was no significant overall subjective preference for engines with male voices over engines with female one's ($t(73)=.687$, $p=.494$), subjects performed better in the synthetic voice condition when listening to a TTS engine with a male voice. Table 5 shows that subjects

in both engine gender groups performed equally well on the comprehension of human speech, but that the subjects who listened to a male synthetic voice performed significantly better on the synthetic section than those who listened to a female voice ($F(1,73) = 9.624$, $p = .003$).

	Human	Synthetic
Male TTS Gp.	72%	65%
Female TTS Gp.	72%	51%

Table 5: comprehension performance by voice type for subjects hearing male and female synthetic voices

Subjective Findings

Regarding subjects' perception of performance, only 16 reported that listening to the TTS was much harder than the human voice, yet scored higher (or equal) in synthetic comprehension than in human. For the majority of subjects (N=49), perception of the increased difficulty of synthetic was correlated to lower performance. Two subjects thought that synthetic speech was actually easier (even though they scored lower with synthetic). The balance (N=11) felt the difficulty was about the same for both voices. A representative comment was "I'm a technophile, but the synthetic speech was too difficult to comprehend. It required much more concentration to listen to synthetic speech than to human speech". Or "I found the synthesized speech rather painful to listen to requiring a great deal of concentration (much like listening to someone with a very heavy accent)."

Even though the human speech was read at a faster word per minute rate than the synthetic (200 wpm vs. 175 wpm), subjects often commented that the speaking rate for TTS was too fast (but did not comment equally on the speed of the human speech). MOS scores for the speaking rate question showed that 36 subjects felt the speaking rate for synthetic was adequate, while 33 felt that it was either a little fast (N=22) or much too fast (N=11). Only one subject thought the rate was too slow (and we believe this subject was thinking more of the process of accessing messages via synthetic speech) and 8 felt the rate was a little slow.

An Evaluation of the Data Sets

As mentioned earlier, we had 10 subjects, independent of those taking the comprehension test, answer all six sets of questions without ever hearing the text. Those subjects answered on average one third of the questions correctly. This result is slightly higher than the 25% that we would have expected given the probability of guessing correctly among the possible 4 choices for each question. When we looked at the results, it was somewhat baffling to us the questions that a majority of these 10 subjects guessed correctly such as "Which three items were you asked to pick up from the store?" a) milk, butter and eggs b) milk, orange juice and butter c) eggs, milk and orange juice d) butter, bread and eggs. The correct answer is c) which does not appear to us any more likely than any of the other choices.

Our goal was to create two sets of data, each consisting of a short, medium and long task, that would be equal in terms of complexity. In reviewing the results by task, it appears that M1 may have been easier than M2. Certainly for the human voice condition, performance scores were the highest for M1 in both the note-taking and non note-taking conditions. (See Tables 6 and 3).

	S1	M1	L1	S2	M2	L2
Hum.	65%	94%	79%	70%	72%	65%
Syn.	52%	77%	48%	75%	77%	70%

Table 6: performance by task for the note-taking condition (N=36)

FUTURE WORK

Something we did not measure in this particular study was the degree to which complex cognitive operations such as decision making or perceptual processing are affected while listening to synthetic speech. If as we suspect, the perception of TTS imposes a greater demand on cognitive resources than perception of human speech, performance on a secondary task would most likely be affected. Given one of the possible scenarios for the application we have in mind, the secondary task could be driving, which could have serious ramifications. Alternatively, a more likely scenario is that the perception of the text-to-speech would decrease due to the concentration effort given to driving. In our future work we want to conduct an experiment that uses similar passages of text, but has the subjects listening to them while in a driving simulator.

Additionally, we would like to run a study that examines the effect of intermingling synthetic speech with human speech. There have been indications in the past that intermingling the voice types decreases comprehension significantly. Because of the need to use synthesized speech for dynamically created text such as e-mail, and the desire to minimize the use of synthetic speech, many systems use a combination of recorded human prompts with synthetically generated speech. It is not clear that we do our users a favor by juxtaposing human speech with synthetic in such systems. We speculate that if we were to run the same test again, the only difference being the intermingling of a synthetic passage with a recorded human passage, the differences in comprehension accuracy between the voice types would be greater.

Another issue that we did not deal with in this experiment, but that needs to be addressed is the pre-processing step that is required for optimal comprehension of synthesized e-mail messages. For our study this was not a problem since we did not include in the messages any headers, footers, imbedded objects or multiple levels of forwarded messages

or replies with history. Markup languages exist for synthetic speech and can be used in a pre-processing step with varying degrees of effectiveness. We would like measure how effectively these conditions can be handled and what the resulting effect on comprehension is.

Lastly we would like to measure, given conditions similar to the study we just ran, how long it would take for naive subjects to achieve comparable comprehension levels between synthetic speech and human, and if this ever happens when listening without notes.

CONCLUSION

While performance for the second set was comparable between synthetic speech and human speech within the note taking condition, it is unlikely that users will have a pad and pen at hand when using pervasive (i.e. ubiquitous) computing devices. One subject commented "I doubt I would recall facts without taking notes and it is doubtful you'd have paper and pencil available all the time (so general comprehension is more germane to the use of this technology)". Given the inability to take notes, we can expect that a majority of users will experience some comprehension problems when listening to messages generated with synthetic speech. Also, given that there was no significant difference in performance across all five engines, these problems will exist regardless of the chosen engine. Since we saw that comprehension difficulties were exarcerbated when the text was longer and more unfamiliar to the subject, we can conclude that a remote access system for messages would have the greatest chance of success if messages were brief and their delivery conducted concisely.

Several people indicated that based on this experience they would not be inclined to use a text-to-speech system for listening to news articles. A representative comment was "Listening to my e-mail would be fine, but not for longer items. I didn't enjoy listening to synthesized news". One of the common culprits mentioned was the difficulty subjects had with hearing and understanding names. Several subjects mentioned that numbers were easily recognizable within the text.

With regard to the potential usefulness of such a system, many subjects commented that such a system would be worth the additional mental effort of listening to synthetic speech for access to critical information such as urgent messages, or for the convenience of rapid access (no computer connections) to e-mail messages. Like any other system, the willingness to submit to the rigors of the system (whatever they be) depends on the value proposition for the end user. This aspect is highlighted in the following comment "I found understanding somewhat difficult, so use of such a system would be highly dependent on what other features/conveniences were included in the system."

ACKNOWLEDGMENTS

Thanks to R. Middlebrooks for invaluable assistance in running subjects as well as in the transcription and analysis of data. Thanks also to J. Karat for his guidance and feedback on the design of the study.

REFERENCES

1. Francis, A.L., Nusbaum, H.C. (1999). Evaluating the quality of Synthetic Speech. In Gardner-Bonneau, D. (Ed.), *Human Factors and Voice Interactive Systems* (pp. 63 - 97)

2. Greenspan, S. L., Nusbaum, H.C., and Pisoni D.B. (1988). Perception of synthetic speech produced by rule: Intellibility of eight text-to-speech systems. *Behavioral Research Methods, Instruments, and Computers*, 18 100-107

3. Klatt, D. H. (1987). Review of text-to-speech conversion for English. *The Journal of the Acoustical Society of America*. September 1987 (pp. 737 - 793)

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

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

6. Ralston, J.V., Pisoni, D.B., Lively, S.E., Greene, B.G. & Mullennix, J.W. (1991). Comprehension of Synthetic Speech Produced by Rule: Word Monitoring and Sentence-by-Sentence Listening Times. *Human Factors* 1991, 33(4) (pp. 471-491)

7. Ralston, J.V., Pisoni, D.B., and Mullennix, J.W. (1995). Perception and comprehension of speech. In Syrdal, A.K., Bennett, R.W., Greenspan, S.L. (Eds.), *Applied Speech Technology* (pp. 233-288) Boca Raton: CRC Press

8. Van Bezooijen, R. & van Heuven, V. (1998) Assessment of Synthesis System. In Gibbon, D., Moore, R. and Winski,R. (Eds.) , *Volume III: Spoken Language System Assessment* (pp. 167[481] - 249 [563]) Mouton de Gruyer

9. www.toefl.org

10. www.genmagic.com/portico/portico_home.shtml

11. www.webley.com