2005

Effect of ambient noise on the comprehension of synthetic speech

Kristen S. Huang
San Jose State University

Follow this and additional works at: https://scholarworks.sjsu.edu/etd_theses

Recommended Citation

Huang, Kristen S., "Effect of ambient noise on the comprehension of synthetic speech" (2005). *Master's Theses*. 2837.

DOI: <https://doi.org/10.31979/etd.xsa3-a3gv>

https://scholarworks.sjsu.edu/etd_theses/2837

This Thesis is brought to you for free and open access by the Master's Theses and Graduate Research at SJSU ScholarWorks. It has been accepted for inclusion in Master's Theses by an authorized administrator of SJSU ScholarWorks. For more information, please contact scholarworks@sjsu.edu.

NOTE TO USERS

This reproduction is the best copy available.

UMI[®]

EFFECT OF AMBIENT NOISE ON THE COMPREHENSION
OF SYNTHETIC SPEECH

A Thesis

Presented to

The Faculty of the Department of
Industrial and Systems Engineering
San Jose State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

by

Kristen S. Huang

December 2005

UMI Number: 1432455

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

UMI Microform 1432455

Copyright 2006 by ProQuest Information and Learning Company.

All rights reserved. This microform edition is protected against unauthorized copying under Title 17, United States Code.

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

© 2005

Kristen S. Huang

ALL RIGHTS RESERVED

APPROVED FOR THE DEPARTMENT OF INDUSTRIAL
AND SYSTEMS ENGINEERING

Dr. Kevin Jordan, Professor

Dr. Mark Van Selst, Associate Professor

Dr. Anthony Andre, Founding Principal, User Interface Analysis

APPROVED FOR THE UNIVERSITY

12/21/05

ABSTRACT

EFFECT OF AMBIENT NOISE ON THE COMPREHENSION
OF SYNTHETIC SPEECH

By Kristen S. Huang

The researcher conducted a study to investigate the effect of ambient noise on the comprehension of synthetic speech in English by native speakers of English, Spanish and Vietnamese. Passages of synthetic speech 300 words in length were presented to participants under quiet and ambient noise listening conditions (60 dB +/-6 dB SPL). An omnibus analysis performed on participants' raw scores for comprehension accuracy revealed a main effect for noise and an insignificant main effect for native language. That is, comprehension scores for all participants were lower in the ambient noise condition regardless of native language. Some findings from this study support the theory that synthetic speech compromises listeners' short-term memory processes. Finally, participants rated the synthetic speech poorly on the portrayal of social characteristics. Conclusions are drawn regarding the application of the results of this study to the design of speech-enabled user-system interactions.

Table of Contents

Introduction.....	1
<i>Prior Research</i>	2
<i>Linguistic Factors</i>	4
<i>Listener and Task Factors</i>	6
<i>Listener Subjective Evaluations</i>	12
Research Questions and Hypotheses.....	14
Method.....	16
<i>Experimental Design</i>	16
<i>Participants</i>	16
<i>Stimuli</i>	18
<i>Apparatus</i>	19
<i>Procedure</i>	20
Results.....	22
<i>Test Passage Difficulty</i>	22
<i>Reading Comprehension</i>	22
<i>Language</i>	23
<i>Listening Environment</i>	24
<i>Subjective Ratings</i>	25
<i>Intelligibility</i>	29
Discussion.....	31
Conclusion.....	37
References.....	39
APPENDIX A: Approval from the Human Subjects-Institutional Review Board...	42
APPENDIX B: Instructions to Participants.....	43
APPENDIX C: Participant Informed Consent.....	44
APPENDIX D: Training/ Test Stimuli and Comprehension Questions.....	46
APPENDIX E: Language Background Questionnaire.....	54
APPENDIX F: Satisfaction Survey.....	55

List of Tables

Table 1. Mean comprehension scores and standard deviations per test passage.....	22
Table 2. Mean RC scores and standard deviations per native language.....	23
Table 3. Means and standard deviations for the omnibus analysis of language.....	23
Table 4. Questions, mean ratings and analysis results for Part I.....	26
Table 5. Responses per language group on part II of satisfaction questionnaire....	28
Table 6. Means and standard deviations for the analysis of presentation order.....	30

List of Figures

Figure 1. Data plot for the omnibus analysis of listening environment.....	24
Figure 2. Means for the omnibus analysis of listening environment.....	24
Figure 3. Means for the analysis of native language by listening environment.....	25
Figure 4: Overall mean ratings on satisfaction questionnaire.....	29
Figure 5: Relative comparison of mean scores and subjective ratings.....	29

Introduction

Advances in technology during the last two decades have made it feasible for people to interact with products and systems using speech. Two types of speech interfaces exist, those that use natural speech and those that use computer generated, or synthetic, speech. Pre-recorded human speech is used to develop natural speech interfaces. The application of natural speech interfaces is limited by two factors: (1) a human voice must be used to digitally record every prompt in the interface, and (2) the resulting file size of the digital recordings is extremely large. Thus, the data storage, transfer and retrieval costs of using natural speech are significant.

Synthetic speech, generated automatically by a software application known as a text-to-speech (TTS) synthesizer, is more practical. A TTS synthesizer reads a simple text file and converts it to spoken language according to a complex set of acoustic and phonetic rules. Depending on the sophistication of the acoustic and phonetic rules, the sound of the synthetic speech can range from mechanical to uncannily human-like.

Speech-enabled interfaces impact the user in many ways. For example, a speech-enabled interface may reduce a system operator's workload in a hands or eyes-free work environment such as an airplane cockpit (Greene, Logan, & Pisoni 1986). Speech-enabled web interfaces can be used to interact with users through portable wireless devices, such as a phone or data assistant, thus allowing people access to information such as Internet content, e-mail, or driving directions. For the visually impaired, a speech-enabled interface can leverage the functionality of an application accessible only through traditional text input/output methods. Synthetic speech-enabled user interfaces

are also easy to update in response to frequent changes, making them ideal for applications that utilize dynamic content, like an e-mail application, a traffic information service, driving directions or an automatic transaction machine.

Speech-enabled user interfaces are becoming prevalent around the world. In fact, in countries where English is the universal language of communication, synthetic speech-enabled user interfaces in English are quite common. As a consequence, non-native speakers of English often use speech-enabled interfaces for travel-related services, such as obtaining flight information in an airport or making reservations. In addition, the technical vocabulary that is used to talk about the World Wide Web and computer terminology in most parts of the world is coined in English. For this reason, research on the comprehension of synthetic speech by non-native speakers of English is vital to the successful deployment of speech technologies in countries where English is used. Despite its utility and sophistication, there are significant processing costs associated with the comprehension of synthetic speech, which could negatively impact the non-native English-speaking population. If non-native speakers of English process synthetic speech differently than native speakers, the system interactions of non-native English speaking users of speech-enabled technologies could be compromised.

Prior Research

Research on the comprehension of synthetic speech over the past several decades falls into three major categories: studies investigating the effect of linguistic factors on the comprehension of synthetic speech, studies that address the influence of listener and task factors on comprehension, and listener subjective evaluations of synthetic speech.

The methods used in past research are varied. Some researchers have employed an instrument known as the Modified Rhyme Test (MRT) (House, Williams, Hecker, & Kryter, 1965). The MRT consists of six 50-word lists of monosyllabic words in English having a consonant-vowel-consonant format (i.e., words like tap, pat, bat), which are synthesized and then presented to participants for recognition. Participants must identify the stimulus word from a list of possible responses (closed format), or provide the missing sounds to words in which the initial and final consonants have been omitted (open format). The MRT is useful in assessing the comprehension of short words as well as the perception of specific phonemes, or basic sounds, of a language.

A method for assessing synthetic speech intelligibility at the sentence-level is an experimental protocol in which both semantically anomalous and semantically meaningful sentences are presented to participants to transcribe. For example, a semantically anomalous sentence in English is “The bitter seed vexes a valid dinner” (Mack, 1988) while a semantically correct sentence is “The bird ate the bitter seed for dinner.” Using semantically anomalous stimuli allows researchers to measure any effect on the perception of meaning of the impoverished signal inherent to synthetic speech. While the comprehension of words and sentences in synthetic speech is well documented in the literature, longer passages of information are used in certain applications of speech technology, such as with an Internet screen reader. This is due to the fact that the sophistication of modern TTS systems allows for more complicated applications of synthetic speech to user-system interaction. Few studies exist, however, that address the comprehension of long passages of information presented in synthetic speech.

Linguistic Factors

Certain linguistic characteristics present in speech, such as prosodic (i.e., changes in pitch and speaking rate), semantic (i.e., meaning) and syntactic (i.e., sentence structure) cues, are crucial to comprehension for many reasons. Wickens (1992) posits that listeners use linguistic cues to figure out the meaning of the message being conveyed. Paris, Thomas, Gilson & Kincaid (2000) add that prosodic cues, in particular, provide important information for comprehension such as the segmentation of acoustic and phonetic units. These researchers also indicate that prosodic cues guide listeners' expectations of semantic information. In this way, prosodic cues serve as a 'load balancing mechanism' for cognitive processing capacity. For non-native speakers of a language, this is particularly important. Depending on the similarity of the phonological rules of a listener's native language to that of the language being used by the system, the presence of well-represented prosodic qualities in synthetic speech can greatly influence how well a non-native listener is able to decipher meaning.

The influence of prosodic cues has been addressed in past studies of synthetic speech intelligibility at the word and sentence level. Greene, et al. (1986) conducted a study to investigate participants' ability to transcribe words on both the open and closed formats of the MRT as a function of synthesizer quality. The investigators compared eight text-to-speech systems ranging in quality from low to high. Results from the study indicated that participants' transcription error rates were lowest for the highest quality TTS synthesizer used. The researchers also found that error rates increased significantly when response constraints were removed in the open format of the MRT and participants

were faced with a more complex transcription task regardless of synthesizer quality. Similar results were obtained in Slowiaczek and Nusbaum (1985). These researchers found that the absence of prosodic cues affected participants' ability to accurately transcribe test sentences when the sentences themselves were syntactically complex. Thus it appears from the results of these studies that the absence of prosodic cues negatively affected the comprehension of synthetic speech as task complexity increased using a TTS synthesizer deemed "high quality" in the 1980's.

Further evidence of the effect of synthesizer quality on the intelligibility of synthetic speech is found in a study conducted by Paris, Gilson, Thomas and Silver (1995). These researchers compared the perception of synthetically produced passages in English using two TTS synthesizers classified with "high intelligibility" and "low intelligibility" to that of digitally recorded natural speech. Results indicated that participants made the greatest number of errors on comprehension tasks when the passages were presented using the "low intelligibility" TTS synthesizer. However, there was no significant difference between comprehension accuracy when participants were presented with speech generated by the TTS synthesizer with "high" intelligibility or the digitally recorded natural speech. These results corroborate those obtained by Greene, et al. (1986) in that the low-quality TTS synthesizers used in both studies posed significant intelligibility problems.

Using more recent speech synthesis technology, Paris, et al. (2000) undertook an investigation into the effect of prosodic cues on the perception of synthetic speech in English. Test stimuli were monotone and inflected semantically meaningful and

anomalous sentences that were generated with a “high-quality” TTS synthesizer. Results indicated that participants were able to correctly transcribe as many words with synthetic as with natural speech when all prosodic cues were present and the sentences were meaningful. Additionally, transcription accuracy was significantly hindered with both natural and synthetic sentences when the sentences were semantically anomalous and presented in the monotone condition.

These parallel results for both natural and synthetic speech indicate that modern TTS technologies are capable of producing high quality speech such that the presence or absence of prosodic cues plays a secondary role to the communication of meaning. Findings from recent studies that reveal more accurate comprehension with modern TTS synthesizers confirm that modern TTS technology is able to mimic the prosodic cues of natural speech with greater accuracy than do older TTS technologies.

Listener and Task Factors

Rabbit (1966) investigated the processing of digits and words presented with natural speech in noise. Rabbit’s results have been cited often in literature on synthetic speech because the degradation in sound quality that occurs with natural speech in noise creates a similar listening environment to hearing the synthetic speech signal. Rabbit found that participants’ ability to recall digits and words was significantly hindered in noise. Rabbit concluded that the proper rehearsal and encoding of the information contained in the speech signal was negatively affected and that participants’ performance decrements were due to the requirement for extra capacity in short-term memory needed to process the speech in noise. Speech is transient and lacks the redundancy of

information that printed text provides. Thus, by nature, speech poses an extra burden to listener's comprehension because it takes more time to process and occupies more capacity in short-term memory due to the lack of redundancy of information.

Pisoni and Koen (1982) investigated the effect of noise on the recognition, identification and understanding of synthetic words using both the open and closed formats of the MRT. Results indicated that words presented in synthetic speech within a specified range of signal to noise ratios negatively affected participants' performance on recognition tasks and that using natural speech under the same conditions did not produce deficits in participants' performance. Thus, in this study the perception of synthetic speech in noise is detrimental to listeners' real-time cognitive processing. A question that has not been addressed in the literature to date is to what extent noise affects the comprehension of longer passages of information, which requires more sophisticated cognitive strategies than the mere recognition of words.

The role of short-term memory in the perception and comprehension of synthetic speech was further addressed in Luce, Feustel and Pisoni (1983). These researchers investigated the effect that a secondary task would have on the recall of synthetic speech. The investigators asked participants to memorize strings of digits presented visually on a screen, in addition to recalling words presented in synthetic speech. Both the recall of digits and words were measured. Results indicated that fewer synthetic words and digits were recalled overall. Most interestingly, the researchers found that when participants were asked to recall the digits and words in the order in which they were presented, digits and words presented at the end of the lists were recalled with greater accuracy. In their

discussion, the researchers point to a failure to rehearse and encode the synthetic words presented earlier in the lists as a cause of the deficits in recall.

Waterworth and Thomas (1985) used a similar methodology to the Luce et al. (1983) study but found that, unlike in Luce et al., digits and words presented at the beginning of the lists were recalled with greater accuracy. Given their contradictory findings, the researchers undertook a subsequent experiment in which participants were informed prior to completing the task that they would be expected to later recall the words and digits. The researchers predicted that by giving participants prior knowledge of the secondary task, that they would utilize more capacity in short-term memory to retain digits and words presented earlier in the lists at the expense of encoding items presented at the end of the lists. Results indicated that indeed, items presented earlier in the lists of both words and digits were recalled with greater accuracy. The results lead the researchers to conclude that synthetic speech is difficult to perceive and understand relative to natural speech because it affects the allocation of limited processing resources in short-term memory. The results also lend some support to the finding of Rabbit (1966) that degraded speech input requires extra capacity in short-term memory, that decrements in recall for degraded speech input result from both encoding and retrieval difficulties.

In a recent study that utilized synthetic speech produced with a modern TTS synthesizer, Lai, Wood and Considine (2000) investigated the comprehension of passages of unrestricted, authentic text commonly found in on-line contexts, like in e-mail messages or news reports. The passages used in the study varied in length from (a) short

(100 words), (b) medium (300 words) to (c) long (500 words). Half of all participants were able to take notes while listening to the passages in synthetic speech.

Results indicated no significant differences on comprehension test scores for natural or synthetic speech. Moreover, comprehension difficulties increased significantly for both types of speech as passages got longer and thus more complex. Those participants who were able to take notes while listening to synthetic speech did perform significantly better than those who did not take notes while listening to synthetic speech. The researchers mention that even though participants did not refer back to their notes, the exercise of taking notes provided them with additional rehearsal of the information in the synthetic passages. The authors posit that participants were able to quickly learn cognitive strategies given the prolonged exposure to synthetic speech that listening to all three passages provided. Finally, the lack of a significant difference between comprehension scores for natural or synthetic speech attests to the fact that modern TTS technology is highly sophisticated in its ability to mimic natural speech.

For non-native speakers of a language, the degraded speech signal inherent to synthetic speech places similar demands on short-term memory processing capacity, as does presenting speech in noise or in conjunction with a secondary task. Mack (1988) investigated the influence of native language on the comprehension of synthetic speech in English. Native language is defined as the language that an individual learns from birth onwards, which is spoken by at least one parent in the home. In her study, groups of native and non-native speakers of English participants were presented with semantically meaningful and anomalous sentences in synthetic speech to transcribe. The non-native

English-speaking participants were all moderately to very fluent speakers of English who resided in the U.S. at the time of the study but had not begun to acquire English until the age of 10 years on average. Results indicated that the non-native English-speaking group made significantly more errors overall than the native English-speaking group, particularly when transcribing anomalous sentences. The author explained that the non-native English-speaking participants lacked the linguistic competence of native speakers of English. Linguistic competence is defined as one's ability to make intuitive judgments about a language in any of the four modalities: listening, speaking, reading or writing (Akmajian, Demers & Harnish, 1984). In other words, the non-native English-speaking participants did not have the intuitive knowledge of English to compensate for the information decrements inherent in the synthetic speech signal, and thus their comprehension of the sentences was hindered.

Evidence for the role of linguistic competence in the comprehension of synthetic speech as a function of one's experience with a language is found in Ozawa and Logan (1989). These investigators presented groups of native and non-native speakers of English participants with words from the open-format of the MRT for transcription. Results indicated that the non-native group made significantly more errors transcribing synthetically produced words than the native group. A closer analysis of the English proficiency level of the non-native speaking group revealed that those participants who self-described their proficiency as good (choosing a 3 or better on a 5 point Likert-style language background questionnaire) performed significantly better than those who described themselves as fair (2 or less). Thus, more experience with the English

language, and consequently more linguistic competence, may have facilitated comprehension for the more competent bilingual participants in this study.

The studies reviewed thus far have focused on the comprehension of words and sentences produced with TTS synthesizers developed in the 1980's and 1990's. In general, the conclusions resulting from these studies confirm that the comprehension of synthetically produced words and sentences is error prone for both native and non-native speakers of a language, and that comprehension difficulty is significantly hindered as task complexity increases. Moreover, the results of studies on the influence of listener factors imply that limitations in short-term memory capacity negatively affect the comprehension of synthetic speech. In these studies, the definition of task complexity has been limited to comprehending synthetic speech while performing a secondary task or processing long passages of speech. However, as Rabbit showed, comprehending speech in noise poses similar challenges to short term memory capacity, as does a secondary task. Given that synthetic speech enabled applications could (and probably will) be utilized in noisy environments, such as over the phone, in noisy places such as an airport or where there is significant noise interference like in an automobile, research that addresses the comprehension of longer passages of synthetic speech in noise merits a closer look.

Particularly interesting are the results gleaned from non-native speakers of English. By nature, this group of individuals has significantly more difficulty comprehending speech in their non-native language even though more exposure to the non-native language may be helpful to some individuals. It follows that the comprehension of synthetic speech for non-native speakers of a language represents an

increase in task complexity over that which would occur for native speakers of a language and similarly, that processing synthetic speech in a non-native language poses extra challenges to the short-term memory capacity of the listener. However, due to the publication dates of Mack (1988) and Ozawa, et al. (1989), the synthesizers used in the studies were not as sophisticated as present-day synthesizers. For this reason further investigation of the comprehension of synthetic speech by non-native listeners using more sophisticated technology could yield different results. Moreover, while there is sufficient evidence that non-native speakers of English have more difficulty comprehending words and sentences in synthetic speech in English, more research is needed to address the comprehension of long passages of synthetic speech by non-native speakers.

Listener Subjective Evaluations

The successful adoption of speech-enabled technologies will depend not only on users' ability to comprehend synthetic speech, but will also hinge on users' satisfaction with using the systems. Evidence exists that humans interact with systems in a manner similar to the way in which they interact with other humans. Reeves and Nass (1999) replicated a variety of classical psychological experiments in which they replaced humans with a computer system in interactions with other humans in order to observe the effect on the interactions. Results of their studies revealed that people exhibit typical social characteristics, such as politeness, when interacting with even the simplest of systems. They propose that the design of system user interfaces should be based on social principles, and add that humans will be more adept at mastering the systems if the

interface displays typical social characteristics. Representing human social characteristics via a user interface is a sensitive endeavor, for the social representation must reflect proper etiquette that can apply to a variety of user populations located in different countries.

Regarding a speech-enabled user interface, Duggan and Deegan (2003) posit that the naturalness of the synthetic voice is an important social characteristic. The more natural sounding the voice, they contend, the more satisfied a user is with the system interaction. Other suggestions they offer include creating a consistent personality for the automated voice that matches the function and task that the user wishes to accomplish with the system. Moreover, the authors caution that a speech-enabled user interface should not offend the user of the system. This is particularly important in inter-cultural contexts where the norms of social interaction are not consistent across cultural groups.

Evidence for the role of social characteristics in conveying a message using synthetic speech is found in Stern, Mullenix, Dyson and Wilson (1999). These researchers investigated participants' perceptions of the persuasiveness of a passage in synthetic and natural speech. Results indicated that there were no statistically significant differences between participant ratings of the persuasiveness of the passage presented in natural or synthetic speech. These findings indicate that participants' perceptions of the persuasiveness of the passage were unaffected by the degraded speech signal inherent to the synthetic speech. Furthermore, these results indicate that social perception is indeed conveyed through synthetic speech, in a manner similar to how it is conveyed through natural speech.

Research Questions and Hypotheses

This study did not compare natural to synthetic speech because results from recent studies, in which modern TTS systems were utilized, have been inconclusive or have indicated no significant differences between the two speech types. Instead, the focus was the comprehension of synthetic speech in both quiet and ambient noise listening environments by non-native speakers of English. The following research questions were addressed in this study: (a) do all participants regardless of native language have more difficulty comprehending unrestricted passages of synthetic speech produced with the highest quality TTS synthesizer in ambient noise? (b) Is one's native language a factor in the comprehension of synthetic speech? And (c), does native language play a role in both native and non-native speakers' of English subjective evaluations of the speech?

Evidence exists that fluency in two languages can have both positive and negative consequences. It has been demonstrated that the age of onset of acquisition of a second language influences the later development of abilities in that language. For example, acquiring a second language at an early age predicts whether a bilingual will possess native-like articulation skills in the second language (Larew, 1961; Asher & Garcia 1969). Evidence also exists that the benefits of early acquisition decline after the age of 10 years (Lenneberg, 1967) and that the presence of a second language can interfere with language processes when the bilingual speakers' two languages are similar (Lambert & Rawlings, 1969). All bilingual participants in this study were *co-ordinate bilinguals* (Baetens-Beardmore, 1986). A co-ordinate bilingual is an individual who has learned a second language some time after infancy but before the age of 10 years, in a setting other

than at home. This group of bilinguals is also commonly referred to as *early* bilinguals (Mack, 1984). The rationale for targeting co-ordinate, or early, bilinguals in this study was to assess whether early acquisition of English was a determining factor in participants' ability to comprehend unrestricted passages of synthetic speech in English produced by a modern TTS synthesizer.

The researcher predicted that all three native language groups would have more difficulty comprehending the synthetic speech passages presented in ambient noise than in the quiet listening environment. Secondly, it is predicted that the native Spanish and Vietnamese-speaking groups would have significantly more difficulty comprehending the synthetic passages than the native English-speaking group in both listening environments. A further prediction involved the structural differences between Spanish and Vietnamese. Vietnamese, like all Asian languages, is a tone language. In a tone language, prosodic elements are used at the syllable level to convey meaning. In Indo-European languages like English and Spanish, prosodic elements like changes in pitch are used to convey meaning at the word and sentence level (Rodman, p. 33). Prosodic elements are thus represented similarly in spoken English and Spanish. An example of the use of prosodic elements in spoken English is when a speaker's voice goes up in pitch at the end of a sentence in a question. The researcher predicted that the close structural relationship between English and Spanish would facilitate comprehension for native speakers of Spanish in one or both listening environments, while there would be no facilitation for native speakers of Vietnamese.

Method

Experimental Design

The study utilized a 2x3 factorial design. The two independent variables were (a) native language and (b) listening environment. The factor of native language had 3 levels (a) English, (b) Spanish and (c) Vietnamese. The factor of listening environment had two levels (a) quiet and (b) ambient noise. The two dependent measures in this study were response accuracy on written tests of listening and reading comprehension and participants' subjective ratings of the synthetic voice

Participants

Forty-eight undergraduate students at San Jose State University were selected to participate in this study. The researcher made initial contact with the students by visiting their Introduction to Psychology courses, and solicited student native language and contact information. The students were then personally contacted by telephone, at which time a language background questionnaire was administered (see Appendix E). Any student whose first language was English, Spanish or Vietnamese was eligible to participate if (a) they were bilingual and learned English as a consequence of studying it in school in the U.S and (b) they rated their ability to comprehend spoken English as a 4 or 5 on the language background questionnaire (on a 5 point scale). Sixteen people speaking each native language of interest were subsequently contacted and an appointment was made with each one to complete the protocol.

Participants ranged in age from 17 to 30 years. Every participant was enrolled in Introduction to Psychology during the fall, 2005 semester at San Jose State University. No participant indicated any hearing problems. No participant indicated having any prior

experience with computer-generated speech. All 16 native English-speaking participants were born in the U.S. and 13 of the native Spanish-speaking participants were born in the U.S.; 3 Spanish-speaking participants were born in Mexico; 8 of the native Vietnamese-speaking participants were born in the U.S.; 7 were born in Vietnam, and 1 was born in the Philippines. All bilingual participants immigrated to the U.S. prior to attending kindergarten. Within each native language group, an attempt was made to achieve a balance of male and female participants. The resulting ratios of male to female participants per language group were 1:1 for the native English-speaking group, 1:3 for the native Spanish-speaking group and 1:1.7 for the native Vietnamese-speaking group. For specific information about participant demographics collected on the Language Background Questionnaire, see Appendix F.

The mean number of years speaking English for the native Spanish and Vietnamese-speaking groups was 13.8 and 12.0 respectively. All 32 bilingual participants stated that they were able to either read or write in their native tongue, with the exception of two native speakers of Vietnamese. All participants indicated that they used their native language at home to communicate with parents and other family members. No native English-speaking participant who spoke a foreign language with close relatives on a regular basis, such that every native English-speaking participant was at least a second-generation citizen of the United States and grew up speaking English in the home.

While fluent in their first languages, native speakers of Vietnamese and Spanish disclosed that they felt most comfortable communicating in English. Moreover, they

disclosed that English was their “inner voice” used when counting money or doing a math problem and rarely, if ever, were able to recall having a dream in their native tongue. This is not surprising, given the age on onset of participants’ acquisition of English (before age 10) and their command of academic English language skills. Every bilingual participant was a university-level student pursuing studies at an institution in which instruction is delivered in English, and it could be said that their linguistic competence in English meets or exceeds that which is required to function in an academic environment. Thus, the individuals who served as participants in this study possessed, at the very least, a level of linguistic competence in English that is consistent with their level of education.

Stimuli

The focus of the present study was the comprehension of unrestricted medium length passages (300 words) of synthetic speech as in Lai et al. (2000). ATT’s state-of-the-art TTS synthesizer, Natural Voices Version 1.4, was used with the American English adult male voice “Mike” to synthesize the passages. The voice of “Mike” was utilized in lieu of the equivalent female voice for two reasons: (a) results from a pilot study in Lai et al. (2000) confirmed that listeners preferred the male voice and (b) because the higher pitch characteristics of the female voice are more difficult to represent with synthesis (Rodman, p. 278), a factor which could potentially impact the results of the study. The passages were saved as .wav files on a PC.

The identical medium length test passages M1 and M2 from the Lai et al. (2000) study were used as stimuli in the present study. The decision to use medium length

passages was based on results from this study. In the study, participants performed better on tests of comprehension of synthetic speech for medium length (300 words) passages than they did for short (100 words) or long (500 words) passages, with a mean comprehension score of 68% for the medium length passages across participants and conditions of task complexity. The mean score of 68% indicates that the passages could be challenging enough to comprehend for native speakers of English, yet comprehensible enough to non-native speakers such that an effect for the independent variables manipulated in this study might be observed. The stimuli were taken from the listening comprehension portion of Barron's Practice Exercises for the TOEFL: Test of English as a Foreign Language (3rd Ed) (1995).

Apparatus

The study took place in a dedicated laboratory set-up in room 335 of Dudley Moorehead Hall on the San Jose State University campus. The room measured 8' in height, was 7.5' deep and 8.5' wide. The ceiling was made of acoustic tile, the walls were made of concrete masonry blocks and the floor covering was linoleum. The apparatus consisted of a table and two chairs, a 40 GB Apple iPod, a Cambridge Soundworks brand Micro Works model amplifier with stereo speakers and a Sony CFD-ZW 700 portable CD/ radio cassette/ player. Stimuli were played with the iPod through the amplifier and speakers at a level of 60 dB +/- 2 dB SPL. Ambient noise was played through the Sony audio device in stereo mode at a level of 60 dB +/- 2 dB SPL utilizing a cassette recording of ambient noise recorded in the food court of a local shopping mall. This sound pressure level is equivalent to the ambient noise present in a crowded

restaurant or conveyed through loud conversational speech, and also represents the minimum sound pressure level that is capable of masking speech (Peterson, 1963). An overall A-weighted average of 60 dB +/- 6 dB was the target SPL in the ambient noise condition, so as to maintain a comfortable listening level as indicated in Rabbit (1966). The SPL level allowed for the speech to be audible, with effort. A few pilot participants were run to assure that the stimuli was audible and a sound meter was used to test the sound pressure level before each test session. Any necessary adjustments were made so that the SPL of each test session was consistent.

The stereo speakers for both audio devices were detached and situated at forty-five degree angles from the participant at an approximate distance of 2.5'. Each individual was tested independently, and during the testing session was seated at the table facing the speakers. Participants were asked to remain facing the speakers throughout the test session. The researcher was seated behind the participant throughout the test protocol. Each test session took approximately thirty minutes to complete.

Procedure

After a short introduction to the study (Appendix B), participants were informed of their rights as participants and were asked to read and sign a consent form (Appendix C). They were advised that any personal information they gave would be kept as confidential. Participants were reminded that they would hear each test passage only one time and would have to recall as many details as possible about the passage on a subsequent test of comprehension for each passage. Every participant heard the same training passage as the first passage presented in the test protocol and were able to self-

check their answers to the comprehension questions accompanying the passage. The presentation of the test passages was completely counterbalanced across groups and listening environments. That is, equal numbers of participants were presented with stimuli in the quiet or ambient noise condition first, while equal numbers were presented with either Test Passage 1 or 2 first. After the second test passage and comprehension test, participants were asked to complete the test of reading comprehension.

Finally, participants rated the quality of the synthetic speech on a 7-point Likert-style post-test questionnaire and were informed to refer to the speech in the quiet condition unless otherwise indicated on the questionnaire. To prevent participants' perceived performance on the comprehension tests from affecting their subjective ratings of the synthetic speech, participants did not have prior knowledge of having to take the satisfaction survey. Rather, participants were asked to complete the survey after they had finished the reading comprehension test. Participants were given the opportunity to take a break after the first test passage, but no participant did so.

Results

Test Passage Difficulty

Comprehension scores were collected during the test protocol for each of Test Passage 1 and Test Passage 2. The comprehension test for each passage was out of 5 points. The overall (n=48) mean comprehension score and standard deviation for Test Passage 1 was 3.35 (1.37) or 70%, and for Test Passage 2 was 3.06 (.89) or 61.2%. Test Passage 2 was slightly more difficult for all participants than Test Passage 1, a finding that is corroborated with results from Lai et al. Participants in the Lai et al. study scored a mean of 77% on Test Passage 1 (M1) and a mean of 72% on Test Passage 2 (M2). The researchers in the Lai et al. study attribute the difficulty of Test Passage 2 to the content of the passage, which consisted of a significant amount of information, including several specific numbers and proportions (see the text for Test Passage 2 in Appendix D). Test Passage 2 was more difficult than Test Passage 1 for each language group as well. The means and standard deviations (in parentheses) per language group are listed in Table 1.

Table 1. Mean comprehension scores and standard deviations per Test Passage.

	Test Passage 1	Test Passage 2
English	3.44 (.81)	2.88 (1.31)
Spanish	3.44 (1.36)	3.37 (0.89)
Vietnamese	3.19 (0.83)	2.94 (1.57)

Reading Comprehension

To determine any pre-existing differences attributable to native language between participants' reading comprehension abilities in English, a reading comprehension (RC) test taken from Barron's Practice Exercises for the TOEFL: Test of English as a Foreign

Language (3rd Ed)(1995) was administered as part of the test protocol. Each participant's RC score was out of a total of 6 points. Mean RC scores, percentage equivalents and standard deviations are listed in Table 2. An ANOVA analysis performed on participants' raw RC scores did not achieve statistical significance ($F [2,45] = 2.13, p > .05$), indicating that there were no pre-existing differences between groups regarding their reading skills in English.

Table 2. Mean RC scores and standard deviations per native language.

	Mean Overall	SD
English	4.50 (75%,)	1.10
Spanish	4.06 (68%)	1.29
Vietnamese	3.5 (58%)	1.67

Language

To test the hypothesis that native language is a factor in the comprehension of passages of synthetic speech in English, an omnibus analysis of native language was performed on participants' raw comprehension scores. Each participant's comprehension score was out of a total of 5 points. Mean comprehension scores, percentage equivalents and standard deviations are listed in Table 3. The 2x3 mixed methods omnibus analysis of variance for the main effect of the factor of language was not statistically significant ($F [2,45] < 1$).

Table 3. Means and standard deviations for the omnibus analysis of language.

	Mean Quiet	SD	Mean Noise	SD
English	3.31 (66%)	1.14	3.0 (60%)	1.10
Spanish	3.94 (79%)	1.10	2.87 (57%)	.96
Vietnamese	3.5 (70%)	1.03	2.63 (53%)	1.31

Listening Environment

To test the hypothesis that all participants would have more difficulty comprehending unrestricted passages of synthetic speech in ambient noise, an omnibus analysis of listening environment was performed on participants' raw comprehension scores in quiet and in noise. The data for the omnibus analysis is plotted in Figure 1. The means and standard deviations were, respectively, 3.58 (1.09) or 72% and 2.83 (1.12) or 57% and are displayed in Figure 2 below. The analysis for the main effect of the factor of listening environment was significant ($F [1, 45] = 13.79, p < .01$).

Figure 1. Data plot for the omnibus analysis of listening environment (n=48)

Figure 2. Means for the omnibus analysis of listening environment (n=48)

To address whether any native language speaking group(s) had significantly more comprehension difficulty in quiet or ambient noise, an analysis was performed to assess the interaction between native language and listening environment. The analysis was not statistically significant ($F [2, 45] = 1.25, p > .05$).

To address whether one of the two bilingual groups had significantly more comprehension difficulty in quiet or ambient noise than the other, an analysis was performed to test the interaction of native language and listening environment for the Spanish and Vietnamese-speaking groups only. This analysis yielded a significant main effect for noise ($F [1, 30] = 16.47, p < .05$), but was not statistically significant with respect to the interaction of native language and listening environment ($F [1, 30] < 1$).

Figure 3. Means for the analysis of native language by listening environment.

Subjective Ratings

On the subjective questionnaire, a poor rating was 4 or less, a neutral rating was between a 4 and 5.5 and a favorable rating was between 5.5 and 7. In percentages, a poor

rating was equivalent to 57% or below, a neutral rating was 57-78% and a favorable rating was 78-100%. To assess if native language was a factor in participants' ratings of the synthetic speech, a one-way ANOVA was performed on the mean ratings per question in Part I between language groups. A significant result was obtained for native language for both question number 7 ($F [2,45] = 3.86, p < .05$) and for question number 11. ($F [2,45] = 3.79, p < .05$). Question text, mean ratings and percentage equivalents per native language and analysis results are listed in Table 4.

Table 4. Questions, mean ratings per language group and analysis results for Part I.

Question	English	Spanish	Vietnamese	F (2,45)	p
Part I:					
1. Was voice pleasant?	3.3 (47%)	3.1 (44%)	4.2 (60 %)	= 2.35	> .05
2. Effort to understand purpose?	3.9 (56%)	3.8 (54 %)	3.9 (56%)	= .05	> .05
3. Was rhythm like human?	2.6 (37%)	2.1 (30%)	2.4 (34 %)	= .82	> .05
4. Was prosody like human?	2.8 (40%)	2.8 (40%)	2.9 (41 %)	= .27	> .05
5. Was voice trustworthy?	4.4 (63%)	3.3 (47%)	3.4 (49 %)	= 1.88	> .05
6. Was voice confident?	4.3 (61%)	3.7 (53 %)	3.9 (56%)	= 1.65	> .05
7. <i>Was length too long?</i>	4.9 (70%)	6.1 (87 %)	4.8 (69 %)	= 3.86	< .05
8. Enthusiastic about Dr. Taylor?	4.1 (58%)	3.8 (54 %)	3.5 (50 %)	= .25	> .05
9. Tuition passage persuasive?	3.9 (56%)	4.0 (57 %)	3.4 (49 %)	= .87	> .05
10. How much effort in quiet?	5.3 (76%)	4.9 (70%)	4.9 (70 %)	= .326	> .05
11. <i>How much effort in noise?</i>	2.7 (38%)	1.8 (26 %)	1.9 (27%)	= 3.79	< .05
12. Will use product with TTS?	3.0 (43 %)	2.5 (36 %)	2.1 (30 %)	= 1.94	> .05

1= [Very Poor], 7= [Very favorable]

Regarding overall ratings of the synthetic speech on social characteristics (Figure 4), participants rated the speech poorly on pleasantness (3.5), trustworthiness (3.7), and

confidence (3.9). Regarding Test Passage 1 about the rising cost of college tuition, participants rated the speech poorly on persuasiveness (3.8). Participants rated the speech in Test Passage 2 as unenthusiastic about introducing the guest speaker (3.8). Participants responded neutrally to a question regarding the amount of effort needed to understand the purpose of the passages (3.9). When asked to compare the rhythm and prosody of the synthetic speech to that of a human voice, participants responded that the rhythm was choppy or mechanical (2.4) and that the speech went up and down in an awkward manner (2.8). Regarding the amount of effort needed to understand the passages in the quiet listening environment, participants responded neutrally (5.0). Finally, participants responded they would be not be inclined to use a consumer product with the synthetic voice as an interface (2.5).

Question text, mean ratings per question and percentage equivalents for Part II are listed in Table 5. Regarding participants' subjective ratings of the volume and pace of the synthetic speech in Part II, the overwhelming majority of participants (98%) found the volume of the synthetic speech in the quiet condition to be "just right," while 81% rated the volume of the speech in noise as "too soft." Sixty-nine percent of all participants said that the pace of the speech in the quiet condition was "just right" while 54% of participants indicated that the pace of the speech in noise was "too fast."

Regarding the researcher's prediction that participants' comprehension scores would corroborate with their ratings of the synthetic speech, there was no corroboration between participants' ratings and their comprehension scores (see Figure 5). That is, regardless of how they felt that they performed on the comprehension tasks, participants

tended to rate the speech poorly. The implications of this finding are mentioned in the concluding remarks of this document.

Table 5. Responses per language group on part II of satisfaction questionnaire.

Part II:	Question	Too Loud/fast? # Responses	Too soft/slow? # Responses	Just right. # Responses
English	Volume in quiet?			16 (100%)
	Volume in ambient noise?	2 (12.5%)	12 (75%)	2 (12.5%)
	Pace in quiet condition?	4 (25%)	1 (6.25%)	11(68.75%)
	Pace in ambient noise?	5 (31.25%)	3 (18.75%)	8 (50%)
Spanish	Volume in quiet?			16 (100%)
	Volume in ambient noise?	3 (18.75%)	11 (68.75%)	2 (12.50%)
	Pace in quiet condition?	1 (6.25%)	1 (6.25%)	14 (87.50%)
	Pace in ambient noise?	11 (68.75%)	2 (12.50%)	3. (18.75%)
Vietnamese	Volume in quiet?	1 (6.25%)		15 (93.75%)
	Volume in ambient noise?		16 (100%)	
	Pace in quiet condition?	7 (43.75%)	1 (6.25%)	8 (50%)
	Pace in ambient noise?	10 (62.5%)		6 (37.5%)

The text of participants' open-ended responses to the question of how to improve the speech to make it more understandable are listed in Appendix F. Interestingly enough, out of the 46 participants who responded to the open-ended question, 8 participants (17%) suggested improving the prosody of the speech, 18 participants (39%) made reference to using a human voice or certain qualities of a human voice, 5 participants (11%) made reference to changes in volume and 14 participants (30%) made reference to a change in the pace of the speech.

Figure 4: Overall mean ratings on the satisfaction questionnaire for all participants.

Figure 5: A relative comparison of mean comprehension scores and subjective ratings.

Intelligibility

To test the hypothesis that participants' comprehension scores ought to improve the longer time they had to "train" on the synthetic voice, an analysis was performed on comprehension scores for the first presentation of stimuli in the quiet listening environment versus scores for the second presentation in the quiet listening environment. The researcher predicted that participants who completed the first task in the quiet listening environment would perform better because they had optimal training conditions,

while participants who completed the second task in quiet would have sub-optimal conditions due to their training having been interrupted by the stimuli presentation in ambient noise. The sub-optimal training conditions would result in lower comprehension scores for these participants. Mean comprehension scores and standard deviations per language for participants who received the quiet condition first versus those who received it second are listed in the Table 6. There was no significant main effect for presentation order ($F [1, 42] = .29, p > .05$) nor a significant main effect for native language ($F [2, 42] = 1.44, p > .05$).

Table 6. Means and standard deviations for the analysis of presentation order.

	Quiet Passage 1 st		Quiet Passage 2 nd	
	Mean	SD	Mean	SD
English	3.13	1.25	3.50	1.07
Spanish	4.25	.89	3.63	1.19
Vietnamese	3.13	.83	3.88	1.13

Discussion

At San Jose State University, a student of English as a second, or foreign, language must achieve a rating of Advanced on the TOEFL in order to be considered for entry into the mainstream academic curriculum, which requires a composite score of 520 out of 677 points or 77% (K. Chavis, personal communication, October 20, 2005). The bilingual participants in the present study earned scores of 3.50 out of 5, or 58% (native-Vietnamese speakers) and 4.6 out of 5, or 68% (native-Spanish speakers), which would classify them as Low Intermediate (57-63%) and Intermediate (63-72%), respectively. A Low Intermediate rating describes a student who is able to manage conversational English and who can write only words, while an Intermediate level student is able to converse about a wider variety of topics in English and can write sentences. In fact, the native English-speaking participants' mean RC score of 4.50 or 75% would qualify them as Low Advanced on the Test of English as a Foreign Language. A Low Advanced rating describes a learner of English who is able to write coherent paragraphs in English but is not yet ready to enter mainstream academic life at a university.

It is necessary to keep in mind that these scores only reflect test results for a small subset of bilingual participants' verbal skills in English. Given that (a) there was no significant difference between the RC scores of the bilingual groups and the RC scores of their English-speaking counterparts ($F [2,45] = 2.13, p > .05$), and (b) all participants were enrolled in a mainstream academic course, it is assumed that the bilingual participants in this study possessed a level of verbal skills in English that would enable them to function as native speakers of English.

For this reason, it is not surprising that the main effect for native language was not statistically significant ($F [2,45] < 1$), and thus neither was the interaction of listening environment and native language ($F [2,45] = 1.25, p > .05$). These results indicate that speaking a native language other than English was not a factor affecting the bilingual participants' ability to perform on the tests of comprehension in this study. Furthermore, the results suggest that the structural similarities between Spanish and English did not facilitate native Spanish-speaking participants' comprehension performance. In fact, results from Ju and Luce (2004) indicate that in order for both of a bilingual's languages to be activated during the perception and recognition of synthetic speech, sounds unique to both languages must be present in the acoustic input.

These results further indicate that the bilingual participants' verbal skills in English are, more likely than not, superior to their verbal skills in their native languages. Additional investigation would be necessary to confirm participants' native language skills, however. A direction for future research in this area would be to focus on the comprehension of passages of synthetic speech by individuals who learned English after the age of 10 years, as in Mack (1988). Perhaps a main effect for language would be observed for this group of bilinguals, who are not considered early acquirers of English.

The significant main effect for listening environment ($F [1,45] = 13.79, p < .01$) indicates that for all participants in this study ($n=48$), the comprehension of 300-word passages of synthetic speech presented in ambient noise, at an overall A-weighted average of 60 dB +/- 6 dB SPL, was more difficult than the comprehension of passages presented in a quiet listening environment (no noise). These results are not surprising,

For this reason, it is not surprising that the main effect for native language was not statistically significant ($F [2,45] < 1$), and thus neither was the interaction of listening environment and native language ($F [2,45] = 1.25, p > .05$). These results indicate that speaking a native language other than English was not a factor affecting the bilingual participants' ability to perform on the tests of comprehension in this study. Furthermore, the results suggest that the structural similarities between Spanish and English did not facilitate native Spanish-speaking participants' comprehension performance. In fact, results from Ju and Luce (2004) indicate that in order for both of a bilingual's languages to be activated during the perception and recognition of synthetic speech, sounds unique to both languages must be present in the acoustic input.

These results further indicate that the bilingual participants' verbal skills in English are, more likely than not, superior to their verbal skills in their native languages. Additional investigation would be necessary to confirm participants' native language skills, however. A direction for future research in this area would be to focus on the comprehension of passages of synthetic speech by individuals who learned English after the age of 10 years, as in Mack (1988). Perhaps a main effect for language would be observed for this group of bilinguals, who are not considered early acquirers of English. The significant main effect for listening environment ($F [1,45] = 13.79, p < .01$) indicates that for all participants in this study ($n=48$), the comprehension of 300-word passages of synthetic speech presented in ambient noise, at an overall A-weighted average of 60 dB $\pm$ 6 dB SPL, was more difficult than the comprehension of passages presented in a quiet listening environment (no noise). These results are not surprising,

Regarding the mean of all participants' (n=48) subjective ratings of the synthetic speech, in Part I of the satisfaction questionnaire, participants tended to rate the synthetic speech poorly, with an average score per question of 3.56 across groups. The reader is reminded that a poor rating is one that is 4 or less, a neutral rating is between a 4 and 5.5 and a favorable rating is between 5.5 and 7. The results from the one-way ANOVA analysis across languages on the mean scores for Question 7., regarding the appropriateness of the length of the passages, was significant for native Spanish speakers ($F [2,45] = 3.86, p < .05$). In other words, native speakers of Spanish felt that the length of the passages was more appropriate than did the native speakers of Vietnamese and English. The one-way ANOVA analysis across language groups on the mean scores for Question 11, regarding amount of effort needed to understand the passage in noise, was significant for native speakers of English ($F [2,45] = 3.79, p < .05$). Native English speakers responded that they put forth significantly less effort (2.7) than the native Spanish (1.8) and Vietnamese-speaking (1.9) speakers to understand the passages.

With regard to subjective ratings, the most significant finding is that participants rated the speech poorly on its ability to represent social characteristics such as persuasiveness, naturalness, confidence and enthusiasm. Moreover, participants responded most frequently that making the speech more "human" was the most frequent response. Evidence exists that matching synthesizer voice and persona to a product's purpose and functionality enhances the user experience (Nass et al., 1999). The fact that participants were not satisfied with the social characteristics of the speech produced by a modern TTS synthesizer indicates that matching a synthetic voice to the function of a

product could pose challenges to the development of product-specific speech-enabled user interfaces. If speech-enabled technologies are to be successfully deployed in user-system interactions for consumer and electronic products, social characteristics ought to be properly represented in the speech. This is especially important for a situation in which more than one product exists that meets user's needs, or multiple input methods exist for a product such that the speech interface is not the only means to interact with the system. Further research would be necessary to investigate to what extent participant subjective ratings affect the adoption of a speech-enabled user interface, as well as which consumer products are appropriate for the application of synthetic speech to user-system interactions.

To address the hypothesis that the processing of synthetic speech overburdens short-term memory resources, changing the protocol to include a different testing format might elicit a main effect for language. In the present study, participants had a 25% chance of guessing the correct answer to each question on the comprehension tests. In fact, the researcher speculates that the ability to guess was greater than 25% for some questions if participants were able to correctly apply the process of elimination to answering the questions. In a future study, the task could be changed so that participants are required to recall specific information in open-ended questions, which would decrease the likelihood that they would be able to correctly guess the answer to any question.

Support for the intelligibility of the TTS system used in this study is found in the lack of a significant effect on participants' performance for amount of training. Participants did not do significantly better given a longer time to train in the quiet

condition. These results suggest that participants may not have had to develop new cognitive strategies to “fill in the information gaps” resulting from the degradation of sound quality in the speech. Thus the TTS system in use, ATT’s Natural Voices Version 1.4, specified the prosodic, syntactic and semantic characteristics of English adequately to the extent that participants were able to apply their already existing cognitive strategies to the comprehension task. Had participants been required to develop new cognitive strategies, a difference in comprehension scores for the second presentation in the quiet condition might exist.

Conclusion

The focus of the present study was the comprehension of passages of synthetic speech in English by native and non-native speakers of English in quiet and ambient noise listening environments. Findings of the present study suggest that the comprehension of 300-word passages of synthetic speech challenged the abilities of all participants, including native speakers of English. Furthermore, the overall significant main effect for noise, achieved in the omnibus analysis as well as in the analysis for bilinguals only, corroborates with the findings of Lai et al. (2000) that the comprehension of passages of synthetic speech is hindered as conditions of task complexity increase.

The findings of this study also suggest that early acquirers of English as a second language function like native speakers of English. In this way, early acquirers of English do not have unique processing needs that set them apart from native speakers, which would impact the design of synthetic speech-enabled interfaces. Given that the non-native English-speaking population forms an ever-growing portion of the market for technological products, developers of speech-enabled systems will be required to meet the processing needs of this population of users. Investigation into the skills of bilinguals who acquired English after the age of 10 years would be necessary to confirm any processing constraints due to native language that affect the comprehension of passages of synthetic speech in English.

The following recommendations are offered for the design of synthetic speech-enabled user interfaces based on the findings of the present study: (1) keep messages conveyed in synthetic speech short, (2) keep comprehension tasks simple, (3) maintain a

volume of at least 60 dB +/- 2 dB if the user interaction takes place in a quiet listening environment (4) limit ambient noise to an overall A-weighted average of *less than* 60 dB +/- 6 dB or increase the volume of the speech in noisy environments and (5) take precaution when applying synthetic speech to user interactions in which the social characteristics of the voice conveyed determine the success of the interaction, as in a bank by phone service where users need to be assured that they have been successful at completing a task.

References

- Akmajian, A., Demers, R. and Harnish, R. (1984). *Linguistics* (2nd Ed). Cambridge, Massachusetts: The MIT Press.
- Asher, J. J. & Garcia, R. (1969). The optimal age to learn a foreign language. *Modern Language Journal*, Vol. 53, pp. 334-341.
- American National Standards Institute (1989). Method for measuring the intelligibility of speech over communication systems (ANSI S3.2-1989—a revision of ANSI S3.2-1960). New York: American Standards Association: Author.
- Baetens-Beardmore, Hugo (1986). *Bilingualism: Basic Principles* (2nd Ed). San Diego, California: College-Hill Press.
- Barron's Practice Exercises for the TOEFL: Test of English as a Foreign Language* (3rd Ed). (1995). Princeton, NJ: Barron's Educational Series Inc.: Author.
- Duggan, Brian & Deegan, Mark (2003). Considerations in the Usage of Text-to-Speech (TTS) in The Creation of Natural Sounding Voice Enabled Web Systems. ACM International Conference Proceeding Series, *Proceedings of the first international symposium on information and communication technologies*, pp. 433-438.
- Environmental Protection Agency (1977). *Behavioral and physiological correlates of varying noise environments*, (No. EPA-6001-77-038). Office of Health and Ecological Effects, Washington, D.C: Author.
- Greene, B., Logan, J. & Pisoni, D. (1986). Perception of Synthetic Speech Produced Automatically by Rule: Intelligibility of Eight Text-to-Speech Systems. *Behavior Research Methods, Instruments and Computers*, Vol. 18, No. 2, pp. 100-107.
- House, Arthur, Williams, Michael, Hecker, H.L. & Kryter, K. D. (1965). Articulation-Testing Methods: Consonantal Differentiation with a Closed-Response Set. *Journal of the Acoustical Society of America*, Vol. 37, No. 1, pp. 158-166.
- Ju, Min & Luce, Paul (2004). Falling on Sensitive Ears: Constraints on Bilingual Lexical Activation. *Psychological Science*, Vol. 15, No. 5, pp. 314-318.
- Lai, J., Wood, D. & Considine, M. (2000). The Effect of Task Conditions on the Comprehensibility of Synthetic Speech. *Proceedings of the Conference on Human Computer Interaction 2000*, pp. 321-328. The Hague: Amsterdam.

- Lambert, W.E. & Rawlings, C. (1969). Bilingual processing of mixed-language associative networks. *Journal of Verbal Learning and Verbal Behavior*, Vol. 8, pp. 604-609.
- Larew, L. A. (1961), Children vs. College Students. *Modern Language*, Vol. 45, No. 1, pp. 22-24.
- Lercher, P, Evans, G. & Meis, M. (2003). Ambient noise and cognitive processes among primary school children. *Environment and Behavior*, Vol. 35, No. 6, pp. 725-735.
- Lenneberg, E. *Biological Foundations of Language*. New York: John Wiley & Sons, 1967.
- Luce, Paul A., Feustel, T. & Pisoni, D. (1983). Capacity Demands in Short-term Memory for Synthetic and Natural Speech. *Human Factors*, Vol. 25, No. 1, pp. 17-32.
- Mack, Molly (1988). Sentence Processing by Non-native Speakers of English: Evidence from the Perception of Natural and Computer-generated Anomalous L2 Sentences. *Journal of Neurolinguistics*, Vol. 3, No. 2, pp. 293-316.
- Mack, Molly (1984). Early bilinguals: How monolingual-like are they? In: Michel Paradis & Yvan Lebrun (eds.), *Early Bilingualism and Child Development*, 161-173. Lisse: Swets & Zeitlinger.
- Ozawa, Kanzaori & Logan, John (1989). Perceptual Evaluation of Two Speech Coding Methods by Native and Non-native Speakers of English. *Computer Speech and Language*, Vol. 3, pp. 53-59.
- Paris, C.R., Gilson, S.D., Thomas, M. H. & Silver, N.C. (1995). Effect of Synthetic Voice Intelligibility on Speech Comprehension. *Human Factors*, Vol. 37, No. 2, pp. 335-340.
- Paris, C.R., Thomas, M H., Gilson, R. & Kincaid, Peter (2000). Linguistic Cues and Memory for Synthetic and Natural Speech. *Human Factors*, Vol. 42, No. 3, 421-431.
- Peterson, Arnold (1963). *Handbook of Noise Measurement*. West Concord, Massachusetts: General Radio Company.
- Pisoni, David B. & Koen, Esti (1982). Some comparisons of intelligibility of synthetic and natural speech at different speech-to-noise ratios. *Journal of the Acoustical Society of America*, Vol. 71, Issue S1, pS94.

- Polkosky, Melanie D & Lewis, James R (2003). Expanding the MOS: Development and Psychometric Evaluation of the MOS-R and MOS-X. *International Journal of Speech Technology*, Vol. 6, pp. 161-182.
- Rabbit, P. (1966) Recognition: Memory for words correctly heard in noise. *Psychonomic Science*, Vol. 6, pp. 383-384.
- Reeves, B. & Nass, C. (1999). *The Media Equation*. New York: Cambridge University Press.
- Rodman, Robert D. (1999). *Computer Speech Technology*. Boston: Artech House.
- Slowiaczek, L.M. & Nusbaum, H.C. Effects of Speech Rate and Pitch Contour on the Perception of Synthetic Speech. *Human Factors*, Vol. 27, pp. 701-712.
- Stern, Steven E., Mullenix, John W., Dyson, Corrie-Lynn & Wilson, Stephen J. (1999). The Persuasiveness of Synthetic Speech versus Human Speech. *Human Factors*, Vol. 41, No.4, pp. 588-595.
- Waterworth, John A. & Thomas, Cathy M. (1985). Why is Synthetic Speech Harder to Remember Than Natural Speech? *CHI Proceedings*, April 1985, pp. 201-206.
- Wickens, Christopher (1992). *Engineering Psychology and Human Performance* (2nd ed.). New York: Harper-Collins.

APPENDIX A: Approval from the Human Subjects-Institutional Review Board

San José State
UNIVERSITY

**Office of the Academic
Vice President**

**Academic Vice President
Graduate Studies and Research**

One Washington Square
San José, CA 95192-0025
Voice: 408-924-2480
Fax: 408-924-2477
E-mail: gradstudies@sjsu.edu
<http://www.sjsu.edu>

To: Kristen Huang
547 S. 12th Street
San Jose, CA 95112

From: Pam Stacks, *Pam Stacks*
Interim AVP, Graduate Studies & Research

Date: May 4, 2005

The Human Subjects-Institutional Review Board has approved your request to use human subjects in the study entitled:

"Effect of Noise on the Comprehension of Synthesized Speech in English by Native Speakers of English, Spanish and Vietnamese."

This approval is contingent upon the subjects participating in your research project being appropriately protected from risk. This includes the protection of the anonymity of the subjects' identity when they participate in your research project, and with regard to all data that may be collected from the subjects. The approval includes continued monitoring of your research by the Board to assure that the subjects are being adequately and properly protected from such risks. If at any time a subject becomes injured or complains of injury, you must notify Pam Stacks, Ph.D. immediately. Injury includes but is not limited to bodily harm, psychological trauma, and release of potentially damaging personal information. This approval for the human subjects portion of your project is in effect for one year, and data collection beyond May 4, 2006 requires an extension request.

Please also be advised that all subjects need to be fully informed and aware that their participation in your research project is voluntary, and that he or she may withdraw from the project at any time. Further, a subject's participation, refusal to participate, or withdrawal will not affect any services that the subject is receiving or will receive at the institution in which the research is being conducted.

If you have any questions, please contact me at (408) 924-2480.

Cc: Dr. Kevin Jordan

The California State University:
Chancellor's Office
Bakersfield, Channel Islands, Chico,
Dominguez Hills, Fresno, Fullerton,
Hayward, Humboldt, Long Beach,
Los Angeles, Maritime Academy,
Monterey Bay, Northridge, Pomona,
Sacramento, San Bernardino, San Diego,
San Francisco, San José, San Luis Obispo,
San Marcos, Sonoma, Stanislaus

APPENDIX B: Instructions to Participants

1. The experiment we are asking you to participate in will help us to understand and evaluate how native speakers of languages other than English comprehend speech that is generated by a computer.
2. You will be asked to listen to three passages of information generated by a speech synthesizer. The first passage will serve as practice for you to familiarize yourself with both the sound of synthetic speech and the type of questions you will be required to answer. The next two passages are considered test passages, which you will listen to under different conditions. After listening to each passage you will be asked to recall as many details that you heard as possible on a short test of comprehension.
3. The entire process will take approximately 30 minutes and you will have the opportunity to take a break after the first test passage. The test session ends after you have completed all tasks.

APPENDIX C: Participant Informed Consent

TITLE: A Study of the Comprehension of Computer Generated Speech in English

by Native Speakers of Spanish and Vietnamese

Kristen S. Huang, Principal Investigator

1. There are no anticipated risks associated with performing these tasks. The benefits of participating in this study are the following: Introduction to Introduction to Psychology students who successfully complete the test protocol will receive one hour of research credit for their course.
2. All data you provide will be treated as confidential. If the test sessions are videotaped, no portion of your face or any personal information will be recorded on the tape. If the test results are published, no personal information of any kind will identify you as a participant. No one other than the investigator will have any knowledge of the answers that you provide on any questionnaire or test during this study.
3. You may quit the study at any time. No service of any kind, to which you are otherwise entitled, will be lost or jeopardized if you choose not to participate. Your relationship with SJSU will not be affected in any way. However, in order for you to receive course credit, you must complete the entire test protocol. The investigator will provide you with a copy of this agreement.
4. Questions about this research project can be directed to the investigator, Kristen S. Huang at 408-298-0721 or to the Department of Psychology,

SJSU: 408.924.5600. Any complaints about this research may be presented to Dr. Bienenfeld, Ph.D. (Chair, Department of Psychology, SJSU: 408.924.5600). Questions or complaints about research, participants' rights, or research-related injury may be presented to Pamela Stacks, Ph.D. (Interim Associate Academic Vice President for Graduate Studies and Research, SJSU: 408.924.2480).

Please sign below to indicate your voluntary consent to participate.

Signature (Participant)

Date

Signature (Investigator)

Date

APPENDIX D: Training/ Test Stimuli and Comprehension Questions

Training Passage:

When I was asked to speak at the Community Book Club luncheon, I thought about several topics that might be of interest to a group of avid readers, and I considered doing a book review or discussing the life of a well known literary figure, but I thought that I would break with tradition today in order to share some anecdotes from the life of a man, who, like you, enjoyed reading. You probably already know a lot about this man and have heard his name mentioned along with other great inventors.

In spite of the fact that Thomas Alva Edison had almost no formal education, spending only three months in school, his mother taught him to read at quite an early age. Between the ages of nine and twelve, he read such difficult volumes as Humes History of England, Gibbon's Decline and Fall of the Roman Empire, and Newton's Principia.

As a young man, Mr. Edison decided to read all of the books in the Detroit public library, systematically, shelf by shelf. After finishing the first fifteen feet, he decided to reconsider the task. A few years later, in Cincinnati, his love for reading almost cost him his life. Having stayed at the library very late, Mr. Edison started for home with a pile of magazines for which he had paid the large sum of two dollars. Suspecting that he might be a thief, a policeman ordered him to stop. But Mr. Edison was too deaf to hear the order. The policeman shot, and missed.

In addition to the electric light, Mr. Edison is known for inventing the phonograph, the microphone, mimeograph, electric storage battery, and photographic

film. Whenever he was paid for an invention, he used the money for his two loves - more experiments and more books.

Questions:

- 1) What is the main subject of the speaker's story?
 - a) The life of great inventors in American history.
 - b) Thomas Alva Edison's education
 - c) The life and times of avid readers
 - d) The life of Thomas Alva Edison.
- 2) How did the speaker describe Thomas' age when he read the difficult volumes such as Humes History of England?
 - a) He was 15
 - b) Between the ages of 9-12
 - c) At an early age
 - d) As a young man
- 3) What happened to Thomas when he was on his way home from the library?
 - a) He got arrested for stealing magazines.
 - b) A policeman shot him in the leg.
 - c) A policeman stopped him for being at the library late.
 - d) A policeman fired a shot at him but missed.
- 4) According to the speaker, how did Thomas use the money he earned?
 - a) To do more experiments and buy more books.
 - b) He saved it to donate to the Detroit Public Library.
 - c) He purchased a microphone, mimeograph, electric storage battery, and photographic film.
 - d) He used it towards his college tuition.
- 5) Which word best describes the tone of this talk?
 - a) Critical
 - b) Persuasive
 - c) Informative
 - a) Defensive

Test Passage 1:

It is a great honor for me to introduce today's guest lecturer, Dr. C. Henry Taylor, a colleague who is so well known to all of us for his many accomplishments and contributions to the field of architecture and planning that it hardly seems necessary to recount them here. Nevertheless, as customary, I will summarize his long experience as a prelude to his address. Let's begin with Dr. Taylor's formal education and some of his accomplishments in the field of higher education.

Dr. Taylor received his B.A. degree in urban history at Yale University in 1955. Five years later, he was awarded a master's degree in architecture and a Ph.D. in urban design from Cornell University. Upon graduating, he accepted a teaching position in the Department of Architecture and Fine Arts at Illinois University, where he was promoted to chairman of the department in 1969. Last year Dr. Taylor resigned from the University in order to accept a research position with the Department of Housing and Urban Development in Washington as director of planning.

In addition to teaching and research, Dr. Taylor has devoted much time to designing, writing, and lecturing. Some of his most famous buildings are right here in the Chicago area, and include the Marina Towers office building and the St. Lawrence Seaway Recreation Center. His many articles on functional architecture and urban planning have appeared in scores of journals over the past twenty years, and a half dozen textbooks are to his credit, including one of the books that we use for this seminar, *Trends in Urban Design*.

Today, Dr. Taylor will speak to us about federal regulations for urban development, and frankly, I know of no one more qualified to address the subject. Ladies and gentlemen, please welcome Dr. C. Henry Taylor, scholar, designer, and author.

Questions:

- 1) What is the topic of Dr. Taylor's lecture?
 - a) Federal regulations for urban development
 - b) Trends in urban design
 - c) Functional architecture
 - d) Urban planning
- 2) From which university did Dr. Taylor receive his Ph.D.?
 - a) Yale University
 - b) Cornell University
 - c) Illinois University
 - d) Washington University
- 3) Where is Dr. Taylor employed now?
 - a) Illinois University
 - b) The Marina Towers office building
 - c) The Department of Housing and Urban development
 - d) As an Editor for a journal for architects
- 4) How does the speaker feel about Dr. Taylor?
 - a) He does not want to introduce him
 - b) He does not know him
 - c) He respects him
 - d) He does not believe that he is qualified
- 5) What will the speaker and audience most probably do next?
 - a) Ask Dr. Taylor to sign a book
 - b) Ask questions of Dr. Taylor
 - c) Listen to Dr. Taylor's speech
 - d) Give Dr. Taylor an award

Test Passage 2:

The cost is going up for just about everything, and college tuition is no exception. According to a nationwide survey published by the College Board's Scholarship Service, tuition at most American universities will be on an average of 9 percent higher this year than last. The increase in tuition is expected to affect students by the end of the academic year.

The biggest increase will occur at private colleges and universities. Private colleges are funded differently than public colleges. Public colleges, which are heavily subsidized by tax funds, will also increase their tuition, but the increase will be a few percentage points lower than private colleges.

As a follow up, the United Press International did their own study at several private colleges. At Ivy League schools, such as Harvard and Yale, most advisors recommended that students have at least \$16,800 available for one year's expenses, including \$11,200 for tuition, \$4,800 for room and board, and \$800 for books and supplies. Ten years ago the tuition was \$5,300. To put that another way, the cost has climbed more than 200 percent in the last decade. An additional burden is placed on out-of-state students who must pay extra charges. In addition, foreign students, who are not eligible for scholarships at state-funded universities, must come up with the money on their own.

On the brighter side, the survey revealed that college graduates are entering the best job market since the middle 1960s. Job offers have increased 16 percent from last year, and salaries are equally good, at least for graduates in technical fields. For

example, a recent graduate in petroleum or computer engineering can expect to make as much as \$40,000 per year. A student with a liberal arts degree might expect to make about half that salary, about \$20,000 per year.

Questions:

- 1) What is the main subject of the speaker's story?
 - a) Increases in tuition and fees at American universities
 - b) The cost of attending Ivy League schools.
 - c) The financial burden for out-of-state residents and foreign students at American universities.
 - d) The job market for recent graduates
- 2) What is the average increase in tuition expenses at American universities this year over last?
 - a) 90%
 - b) 20 %
 - c) 9%
 - d) 200 %
- 3) How much did tuition increase at Ivy League schools over a ten-year period?
 - a) \$5,300
 - b) 2%
 - c) \$16,800
 - d) 200%
- 4) According to the reporter, what is a problem for foreign students at state universities?
 - a) They pay more than out-of-state students
 - b) Only a few are accepted
 - c) They are not eligible for scholarships at state universities
 - d) Their scholarships are very small
- 5) Which word best describes the tone of this talk?
 - d) Critical
 - e) Persuasive
 - f) Informative
 - g) Defensive

Test 3: Reading Comprehension in English

Edwin Hubble was an American astronomer whose research led to discoveries about the nature of the universe. He settled a long debate by demonstrating that the Andromeda nebula did reside outside our galaxy. This discovery led to the establishment of the islands universe theory, which states that galaxies exist on their own. His study on the distribution of galaxies resulted in Hubble's Constant, a standard that computes a galaxy's distance from the earth and its speed of recession.

By 1925, Hubble had devised a classification system for the structure of galaxies that provided conclusive observational evidence for the expansion of the universe. His work pushed the Mount Wilson telescope beyond its capability and provided strong impetus for the construction of an instrument twice its size at Mount Palomar, which Hubble used during his last years of life. The telescope that bears his name was launched on a space shuttle in 1990 and orbits the earth gathering information about the size and age of the universe.

1. With what topic is the passage primarily concerned?
 - a. The Hubble telescope
 - b. The nature of the universe
 - c. Edwin Hubble's research
 - d. Hubble's classification system for the galaxies
2. Hubbles' Constant states
 - a. That galaxies exist outside our galaxy
 - b. That there is a fixed relationship between the distance of a galaxy from the earth and its speed of recession
 - c. That the universe is expanding
 - d. That the Andromeda nebula is located outside our galaxy

3. The author mentions all of the following as accomplishments of Hubble EXCEPT
- a. Evidence for the expansion of the universe
 - b. A classification system for the structure of galaxies
 - c. The islands universe theory
 - d. A design for the space shuttle
4. The word “capability” in line 3 of the second paragraph refers to
- a. Confidence
 - b. Capacity
 - c. Quality
 - d. Category
5. The word “its” in line 4 of the second paragraph refers to
- a. Hubble’s research
 - b. An instrument
 - c. The Mount Wilson telescope
 - d. The construction of a new telescope
6. The paragraph that would follow these paragraphs probably discusses
- a. Hubble’s life
 - b. The telescope at Mount Palomar
 - c. Recent information about the nature of the universe
 - d. The space shuttle launch

APPENDIX E: Language Background Questionnaire

Note: Answers are listed in italics. Numbers followed by an S refer to native Spanish-speaking participants. Numbers followed by a V refer to native Vietnamese-speaking participants.

Please answer the following questions:

1. How old are you? *17-30 years of age*
2. Where were you born?
13-S were born in the U.S., 3-S were born in Mexico; 8-V were born in the U.S., 7-V were born in Vietnam, 1-V was born in the Philippines; all 16 native English speakers were born in the U.S.
3. What is the first language that you began speaking (if English, skip to Ques.12)?
(16) English (16) Vietnamese (16) Spanish
4. How old were you when you first started speaking *English*?
3-10 years of age
5. Mother's native language: *(16) Vietnamese (16) Spanish*
6. Father's native language: *(16) Vietnamese (16) Spanish*
7. Rate yourself on your ability to understand spoken English on a scale of 1-5:
All bilingual participants rated themselves as a 5, with the exception of one female Spanish-speaking participant who was born in Mexico, and one Vietnamese-speaking participant, born in Vietnam.
8. Between Vietnamese OR Spanish, and English, what language do you feel most comfortable speaking? *All bilingual participants indicated English;*
9. Can you read and write in your first language?
All bilingual participants indicated YES.
10. What language do you *count* in? *All bilingual participants indicated English.*
11. What language do you *dream* in? *All bilingual participants indicated English.*
12. What experience, if any, do you have with computer-generated speech?
No participant indicated any prior experience with synthetic speech.
13. Do you have any hearing problems? *No participant indicated a hearing problem.*

APPENDIX F: Satisfaction Survey

Note: Overall averages (n=48) are indicated after each question.

Part I: Please circle the number that represents your response to each question below.

1. Was the voice you heard pleasing to listen to? 3.5

1	2	3	4	5	6	7
Very unpleasant.						Very pleasant.

2. How much effort was needed to understand the purpose of the passages? 3.9

1	2	3	4	5	6	7
Very much effort.						No effort required.

3. Did the rhythm of the speech sound like a human voice? 2.4

1	2	3	4	5	6	7
Rhythm was choppy or mechanical.						Rhythm was natural like a human.

4. To what extent did the speech go up and down like a human voice? 2.8

1	2	3	4	5	6	7
Speech went up and down in very awkward manner.						Totally like a human voice.

5. Did the voice appear to be trustworthy? 3.7

1	2	3	4	5	6	7
Not at all trustworthy.						Very trustworthy.

6. Did the voice portray a confident speaker? 3.9

1	2	3	4	5	6	7
Not at all confident.						Very confident.

7. Was the length of the passages appropriate or were they too long for you to be able to understand them? 5.3

1	2	3	4	5	6	7
Too long.						Length not a problem.

8. Recall the passage about Dr. C. Henry Taylor. Did the speaker appear to be enthusiastic about introducing Dr. Taylor? 3.8

1	2	3	4	5	6	7
Not at all enthusiastic.						Very enthusiastic.

9. Recall the passage about college tuition. Was the passage persuasive? 3.8

1	2	3	4	5	6	7
Not at all persuasive.						Very persuasive.

10. Recall the passage that you heard *in the quiet condition*. How much effort did you have to make to understand the passage? 5.0

1	2	3	4	5	6	7
Much effort.						No effort

11. Recall the passage that you heard *in noise*. How much effort did you have to make to understand the passage? 2.1

1	2	3	4	5	6	7
Much effort.						No effort

12. Would you use a product with this voice as an interface? 2.5

1	2	3	4	5	6	7
Very unlikely.						Very likely.

Part II: Please answer the following questions:

1. Was the volume too loud or too soft, or just right in the quiet condition?
47 responded 'just right'; 1 responded 'too loud';
2. Was the volume too loud or too soft, or just right in the noisy condition?
5 responded 'too loud', 39 responded 'too soft' and 4 responded 'just right'
3. Was the speech too fast, too slow or just right in the quiet condition?
12 responded 'too fast', 3 responded 'too slow' and 33 responded 'just right'
4. Was the speech too fast, too slow or just right in the noisy condition?
26 responded 'too fast', 5 responded 'too slow' and 17 responded 'just right'
5. Please describe how you would change the speech to make it more understandable.

In the noisy condition it needs to be louder, the other was ok.

Slower and louder in the noisy condition.

Try to make it less choppy.

Space out words more, it was hard to distinguish the words, jumbled.

I wouldn't change speech, it was understandable.

More like a human voice not a machine.

It was understandable just not that interesting.

Make it sound less like a robot.

I don't know how to make it more understandable.

Having a real human speaking would help a lot more.

Just the noise was a little loud.

Change the voice, it sounds boring.

Change the Speaker's tone of voice.

Too much information was given which made it difficult.

The speech is choppy, too soft in the noise condition.

Use a human voice not a computer.

Make it smoother, not choppy, just like a human's voice.

Sounds too much like a machine, should sound more human.

Too fast, words were skipped over, had to complete words in head.

Make it pause in the right places, words aren't recognizable.

Change to how a regular speaker would speak.

Increase the intonation and make sound more soothing to ear.

The tone is kind of "flat" and it breaks my concentration.

Make it sound more human.

Make it less choppy and change the tone to resemble natural speech.

More like a human voice with feelings and emotions.

More clear and slow.

Speak slower.

*Sound more natural because robotic speech made him easy to tune out.
Be more human and more exciting about a better subject.
It was pretty much understandable.
Make it less choppy and more resembling a human voice, flow was off.
The noise was a distraction but everything else was easy to hear.
Make the tone less flat.
Provide gaps between words.
It didn't seem to fluctuate naturally, but was still understandable.
Use a human.
Careful having two words following each other too closely.
It's too slow, it's hard to comprehend, I lost interest fast.
Have the rates of the voice in noise and quiet be the same.
Put more excitement in the voice, too monotone, speak slower.
I would make the voice speak at a slower pace.
I would insert more gaps between the words to make it more understandable.
I would make the speech less choppy and more natural sounding.
Raise the volume and slow the speech.
Change the tone or try a woman's voice.*