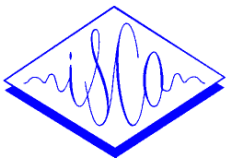

3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

COMPREHENSION OF KTH TEXT-TO-SPEECH WITH "LISTENING SPEED" PARADIGM

Lennart Neovius and Parimala Raghavendra*

Department of Speech Communication and Music Acoustics, KTH,
Stockholm, Sweden
** Names in alphabetic order*

ABSTRACT

The comprehension of natural and synthetic speech in Swedish and American English was investigated using a sentence-by-sentence listening paradigm. The synthesised speech was generated by the KTH text-to-speech systems. Results indicated that sentence listening times were significantly longer only for American English synthetic speech than natural speech. Text difficulty was found to be a significant variable in both Swedish and American English for sentence listening times and word recognition, and only in American English for proposition recognition. The results are discussed in terms of the quality of the synthesisers and factors involved in comprehension.

Keywords: *comprehension, synthetic speech, text-to-speech*

10.21437/Eurospeech.1993-380

1. INTRODUCTION

The extensive application of synthetic speech in aids for the disabled and information technology urge us to investigate how synthetic speech is perceived and understood by listeners. The evaluation procedures for text-to-speech systems are moving from tests that essentially measure aspects of segmental intelligibility [1] to tests that measure comprehension, reading speed and comfort.

Ralston, Pisoni, Lively, Greene and Mullenix [2] investigated the comprehension of natural speech, and synthetic speech generated by a Votrax synthesiser using an on-line task. In this task, subjects listened to 10 different passages, sentence-by-sentence, and they initiated the presentation of each sentence by pressing a button. The time between the end of one sentence and the button command for the next sentence was recorded and this was the sentence-by-sentence listening time (SLT). The underlying assumption of this procedure was that SLTs increased with increasing text difficulty and decreasing speech quality. Subjects were asked to answer whether a test word actually occurred in the passage and whether a proposition in relation to the passage was correct or wrong, after each passage. Five passages were easy (4-grade level) and the other five were difficult (college level). The results showed that SLTs were longer and recognition memory accuracy poorer for passages in synthetic speech than for natural speech. They concluded that on-line comprehension of synthetic speech was poorer than comprehension of natural speech and SLT appeared to be a sensitive measure to study real-time comprehension.

In a pilot study, Carlson, Granström, Neovius and Nord [3] adapted the above paradigm by translating the Ralston et al. [2] text-passages and the questions to Swedish. Thirty-three native speakers of Swedish listened to a practice passage in synthetic speech and two test passages, one in natural speech and one in synthetic speech produced by the KTH text-to-speech research system [4]. The results were different from the previous study. In the study by Carlson et al. [3], the overall SLTs were longer (1 - 1.5 sec vs. .46 - .67 sec), SLTs were the same for easy and difficult passages in synthetic speech, SLTs were longer for difficult text in natural than synthetic speech. Subjects listening to synthetic speech had higher word and proposition recognition accuracy scores than subjects listening to natural speech (89%-95% vs. 78%-88%). Carlson et al. [3] suggested that the differences in results could be due to the use of the higher quality KTH text-to-speech system, the speaking rate, and choice of subjects. Methodological differences such as within-subject design instead of between-subject design, and only oral instructions could also have contributed to different results.

The aim of the current project was to evaluate the comprehension of Swedish and American English KTH text-to-speech research synthesisers using the sentence-by-sentence listening paradigm used by Ralston et al. [2] Our objective was also to investigate whether the sentence listening paradigm was sensitive enough to measure differences between natural, and synthetic speech of various quality and synthesisers in different languages.

2. METHOD

Subjects: Thirty adult native speakers of American English between the ages of 19 and 47, and 40 adult native speakers of Swedish between the ages of 21 and 35 were the subjects. The American English subjects were recruited from Stockholm. The native speakers of Swedish were mostly students of electrical engineering attending a speech communication class.

Stimuli: The text passages contained high information load with new information in every sentence. The easy passages consisted of topics such as birds and bears, and they contained a mean of 10 words per sentence. The difficult passages consisted of topics such as radioactivity and the origin of language, and they contained a mean of 23 words per sentence. The text material and the natural speech files for American English were the same as those used by Ralston et al. [2]. The Swedish passages and the speech files, both natural and synthetic, were the same as those used by Carlson et al. [3].

Instrumentation: The experiments were run on an HP/Apollo workstation. Playback of sentences was administered via a signal processor card, with a D/A converter, which also had the trigger button connected to it. This alleviated the problem of precise time keeping on the UNIX system.

Procedure: The 30 American English subjects listened to 10 passages of text; 15 subjects listened to the passages in natural speech and the other half listened to the same passages in synthetic speech. The 40 Swedish subjects listened to five passages of text; 20 subjects listened to five passages (3 easy and 2 difficult) either in natural or synthetic speech and the other half listened to the other five passages (2 easy and 3 difficult) either in natural or synthetic speech.

Subjects in both language groups were randomly assigned to natural or synthetic speech and the passages were also presented in a random order. The SLT, word and proposition accuracy scores were recorded. Both groups used the same instrumentation and received the same set of instructions on the computer screen and through headphones. Subjects were instructed that they would be in control of the presentation of each sentence in a story. They were instructed that they could spend as much time as they wanted on each sentence, but to keep their finger on the button at all times so that as soon as they understood the sentence, they should proceed to the next by pressing the button. They were also explicitly asked to understand the information presented as they had to answer some questions at the end. After instructions, the subjects listened to one practice passage followed by test passages. They answered 4 questions related to word recognition and 4 questions related to proposition recognition at the end of each passage.

3. RESULTS

The results are discussed in terms of training effect, sentence-by-sentence listening times and word and proposition recognition scores. As in [3], SLTs longer than 4 seconds were discarded before analysing the data as subjects might have taken a break or simply forgotten to press the button.

The data for Swedish subjects listening either to 2 easy or 3 easy and 3 difficult or 2 difficult passages either in natural or synthetic speech were compared using *t*-tests for independent samples using SPSS for PC. It was found that there was no significant difference between subjects listening either to 2 easy or 3 easy and 3 difficult or 2 difficult passages either in natural or synthetic speech. Hence, these data were collapsed to arrive at scores for 10 passages for 20 subjects.

The data were analysed using Analysis of Variance (ANOVAs) [5] with type of speech (natural or synthetic) as a between subject factor and text difficulty (easy or difficult) and word and proposition recognition accuracy as within- subject variables. In order to reduce variability, SLTs greater or less than two standard deviations from the mean for easy and difficult texts for both voices were replaced with the mean value for the same groups based on the procedure used by Ralston et al. [2]. Then SLTs for each subject was averaged across sentences within a passage and across easy and difficult passages.

3.1 Training effect

Figure 1 shows the SLTs for the practice passage and the corresponding regression lines. The individual variation among different sentences reported in [2] and [3] was also found in this experiment. The peak SLT for sentence 16 was due to shift of focus in the story. There was a noticeable training effect for synthetic speech in both languages. The mean SLTs for the first six to seven sentences in this category was above 1.0 seconds. However, the SLTs were reduced to approximately .8 seconds for the later sentences. The slope of the regression lines were also larger for synthetic than for natural speech. The correlation was .62 for Swedish and .61 for American English. The same effect was not seen for natural speech. The first sentence seemed to always have a longer response time, probably due to identifying the actual topic, but the SLTs for later sentences seem to level off. The correlation was .52 for Swedish and .18 for American English natural speech.

3.2 Sentence-by-sentence listening time

Figure 2 shows the overall SLTs for Swedish and American English subjects. SLTs were significantly longer for synthetic speech than for natural speech in American English, $F(1,28) = 5.75$, $p < .025$. However, SLT data for subjects listening to the Swedish synthetic speech were not significantly different from subjects listening to natural speech. Subjects in both language groups

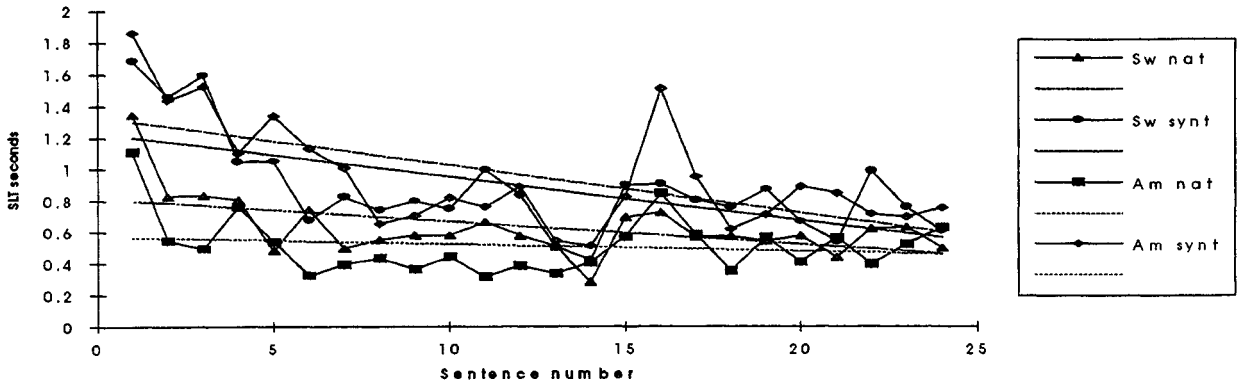

Figure 1. The mean sentence-by-sentence listening times for the practice passages and the corresponding regression lines.

took longer time to listen to difficult texts than easy texts [Swedish, F (1,18) = 8.56, $p < .01$; American English, F (1,28) = 28.28 $p < .001$]. The results for American English were similar to Ralston et al.'s findings even though the listening times were higher overall in the present study. The lack of significant difference between the two groups of Swedish subjects could be due to the higher quality Swedish synthesis. This might have helped the Swedish subjects listening to synthetic speech to perform similar to subjects listening to natural speech. This group was also more homogenous in terms of age range and their background.

3.3 Word recognition

Figure 3 shows the word recognition scores following natural and synthetic speech for easy and difficult passages in Swedish and American English (82-92%). Interestingly, subjects in both language groups did not show a significant difference in their word recognition following natural or synthetic speech. However, text difficulty seemed to have had a significant effect for both the language groups [Swedish, F (1,18) = 4.44 $p < .05$; American English, F (1,28) = 9.03 $p < .01$]. There was no significant interaction effect between voice and text difficulty. The mean values indicated that the Swedish subjects had better recognition accuracy scores for easy text than difficult text. However, American English subjects had better recognition accuracy scores for difficult passages than easy passages. The Swedish subjects behaved similar to Ralston et al.'s subjects. However, subjects listening to American English difficult texts either in natural or synthetic speech seemed to have "paid more attention" to the words in these texts, thus obtaining higher scores.

3.4 Proposition recognition

Figure 4 shows the proposition recognition scores for natural and synthetic speech, for easy and difficult passages in Swedish and American English. A striking feature is that these scores are much higher overall than word recognition scores (91%-97%). Subjects answering questions related to content following natural or synthetic speech, either for easy or difficult passages in Swedish performed in the same manner. However, American English subjects showed a significant difference in proposition recognition accuracy related to text difficulty, F (1,28) = 5.40 $p < .05$. There was some significant interaction between voice and text difficulty, F (1,28) = 4.21 $p < .05$. As

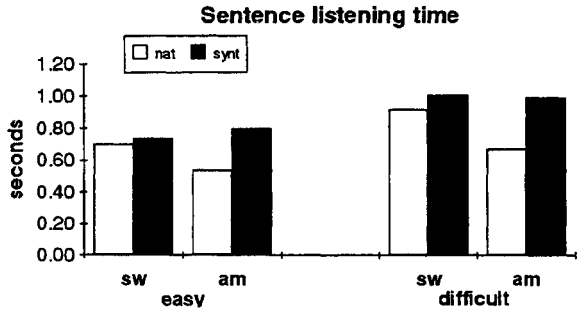

Figure 2. The mean sentence-by-sentence listening times for the test passages in Swedish (sw) and American English (am).

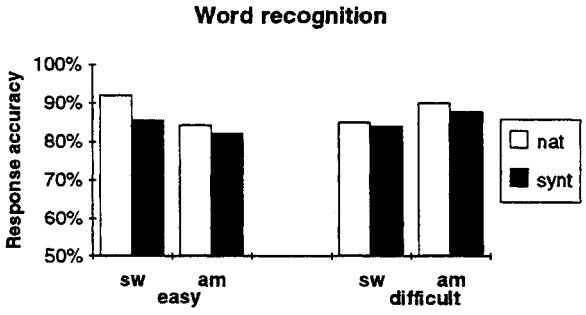

Figure 3. Mean response accuracy for word recognition in Swedish (sw) and American English (am).

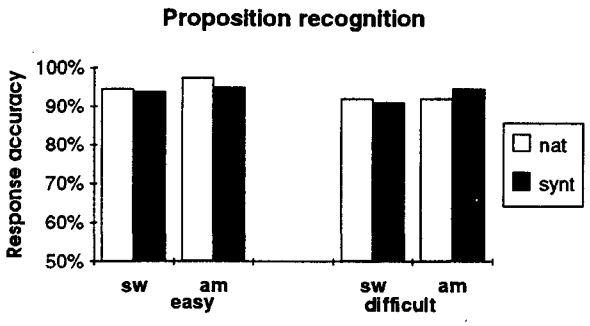

Figure 4. Mean response accuracy for proposition recognition in Swedish (sw) and American English (am).

can be seen in Figure 4, subjects listening to natural speech did better on easy passages than difficult passages, whereas subjects in synthetic speech group performed better on difficult passages than easy passages. A possible explanation could be that subjects listening to difficult text in synthetic speech listened longer and more carefully to each sentence than the subjects in natural speech, and were able to verify the information more accurately, a trade-off between speed and accuracy.

In general, subjects commented that the task was fun and that they "learned a lot", indicating that they had understood the information presented.

4. DISCUSSION

This study investigated the comprehension of natural and synthetic speech in Swedish and American English using an on-line task. Subjects in the two different language groups performed similarly on some aspects and differently on others. A major area of difference was in SLTs. The lack of significant difference between subjects listening to synthetic or natural speech in Swedish could be due to the higher quality of the Swedish KTH text-to-speech system. The longer SLTs for American English KTH system suggest that comprehension proceeded more slowly for passages of synthetic speech than natural speech reflecting on the quality of the synthesiser. The methodology was the same for both language groups except for the number of passages and the translated material. However, we do not consider this to have contributed to the different results.

The higher overall proposition recognition scores reflect that subjects listening to Swedish and American English synthesisers performed well even though the American English subjects required longer time to understand the sentences. The trade-off between speed and accuracy appears to work only with higher quality KTH synthesisers as Ralston et al. [2] subjects listening to Votrax performed significantly poorly on recognition accuracy (72%-80%). The self-paced nature of the task appears to enhance comprehension with longer SLTs and allows subjects to seek optimal responses.

5. CONCLUSIONS

From the above investigation, it is seen that this paradigm reliably differentiates the effect of text-difficulty. Subjects performed better listening to easy texts than difficult texts on most variables. However, the paradigm may not be sensitive enough to measure differences between natural and synthetic speech when higher quality synthesisers, such as the Swedish KTH text-to-speech system, are used. One can also argue that performance on higher quality synthesis is getting closer to performance on natural speech, thus making it difficult to measure significant differences in experimental situations. For American English, subjects took longer time to comprehend the synthetic than the natural speech passages. However, the lack of significant difference for word and proposition recognition scores following natural or synthetic speech shows that the subjects in both groups comprehended the information similarly. There is further need to explore other paradigms that are sensitive to demonstrate differences between natural and synthetic speech of various quality.

6. ACKNOWLEDGEMENTS

We thank David Pisoni and Scott Lively for providing us with the materials and Rolf Carlson and Björn Granström for their support and encouragement. This work was supported by grants from Swedish Language Technology Program and KTH.

REFERENCES

[1] Carlson, R., Granström, B. & Nord, L.: "Segmental evaluation using the ESPRIT/SAM test procedures and monosyllabic words", *Talking machines: Theories, Models, and Applications* (G. Bailly & C. Benoit, ed.'s), 1992a.

[2] Ralston, R., Pisoni, D., Lively, S., Greene, B., & Mullenix, J.: "Comprehension of synthetic speech produced by rule: word monitoring and sentence-by-sentence listening times.", *Human Factors,* 33(4), pp. 471-491, 1991.

[3] Carlson, R., Granström, B., Neovius, L., & Nord, L.: "The "listening speed" paradigm for synthesis evaluation.", *Papers from the Sixth Swedish Phonetics Conference,* Gothenburg, pp. 63-66, 1992b.

[4] Carlson, R., Granström, B. & Hunnicutt, S.: "Multilingual text-to-speech development and applications", in A.W. Ainsworth (ed.), *Advances in speech, hearing, and language processing,* JAI Press, London, 1990.

[5] Brunning, J. L. & Kintz, B. L.: "Computational Handbook of Statistics", Scott, Foresman and Co., Illinois, 1968.