THE "LISTENING SPEED" PARADIGM FOR SYNTHESIS EVALUATION

Rolf Carlson, Björn Granström, Lennart Neovius and Lennart Nord[*]

Department of Speech Communication and Music Acoustics

KTH, Box 70014, 10044 Stockholm

INTRODUCTION

In the current work we have the ambition to expand our test procedures for text-to-speech systems, from tests that essentially measures aspects of segmental intelligibility (Carlson et al. 1992) to tests, that tax the cognitive ability of speech understanding.

In the present study we have adopted a new paradigm introduced by Ralston, Pisoni, Lively, Green and Mullennix (1990) to Swedish. Passages are presented sentence by sentence to one subject in either synthetic or natural speech. The subject controls the playback of each sentence, by pressing a button. The time between the end of one sentence and the button command for the next sentence is recorded. According to the paradigm this time is expected to vary with factors such as speech quality and text complexity. As part of the task the subjects are asked to answer two types of questions after each passage. The subject is not aware of the recording of the time events. However, the subject is specifically instructed to answer to the questions:

- If a test word actually occurred in the passage
- If a proposition, in relation to the passage, was correct or wrong

Both types of questions are easy to evaluate (right/wrong) and has a direct relation to the retention/comprehension of the passage.

In the original study 11 text passages were used. Six were quite simple ("fourth grade level") and five were more advanced ("college level"). One of the simple passages ("Joker") was used for practice. The material was (freely) translated, along with the questions on word occurrence and propositions. The easy passages contained a mean of 9.92 words/sentence and the difficult, 19.3 words/sentence.

In the Ralston et al. study, the synthesis used was produced by the Votrax synthesizer, a widely used device with relatively low quality, as judged from previous studies (Logan, Greene & Pisoni, 1989). All three test measures, sentence listening time, word recognition and proposition recognition, correlated with text difficulty and speech quality. The listening time was especially discriminating.

THE SWEDISH TEST

The Swedish version of the test has been tried in a pilot experiment, with the ambition to get a first evaluation of the test material and the procedures. Thirty-three students of electrical engineering, attending the speech communication class, served as subjects. The playback of the sentences was administered by a signal processor card with D/A converter, installed in an Apollo/HP workstation. This alleviated the problem of precise time keeping on the UNIX system. The button that initiated the playback of new sentences was connected to the signal processing card. Word and proposition recognition was recorded on the UNIX machine.

---

[*] Names in alphabetic order

Each session started with one of the synthetic-speech passages, "Joker", as practice of the procedure and to give some initial exposure to synthetic speech. Each subject was given two more passages, one in natural speech and one synthetic. The synthesis used was produced by the KTH text-to-speech system (Carlson, Granström & Hunnicutt, 1990). The test followed a rotated design.

Sentence-by-sentence listening time

The sentence-by-sentence listening time (SBSLT) is defined in the paradigm to be the time between the end of a sentence and the response by the subject in terms of pressing the continue button.

Practice passage

The subject responses were analysed in terms of the sentence-by-sentence listening time (SBSLT) and the correct score on the word and proposition recognition task. Looking at the raw data, the SBSLT task introduces an analysis problem. It is in fact possible to get values shorter than a basic reaction time, even negative values, i.e. subjects requesting the next sentence before the first is finished. We discarded such data by setting a lower limit of 0.1 sec. Also occasional very long durations were recorded. It seemed that subjects took a break during passages or simply forgot to press the button. Recovering from such an error could take several seconds. Thus, we discarded data in excess of 4 seconds. All data presented below are means from observations in the interval 0.1 sec. to 4 sec.

The SBSLT for the practice passage is presented in Figure 1. As can be seen there is a mild training effect (decrease in the SBSLT) for the first half of the passage. The second part of the curve seems to level off at about 1.1 sec. A regression line is included in the figure with a correlation of .62. Individual variations for the different sentences could be seen, and are also expected due to differences in e.g. complexity of the sentences. This was also observed in the Ralston et al. study.

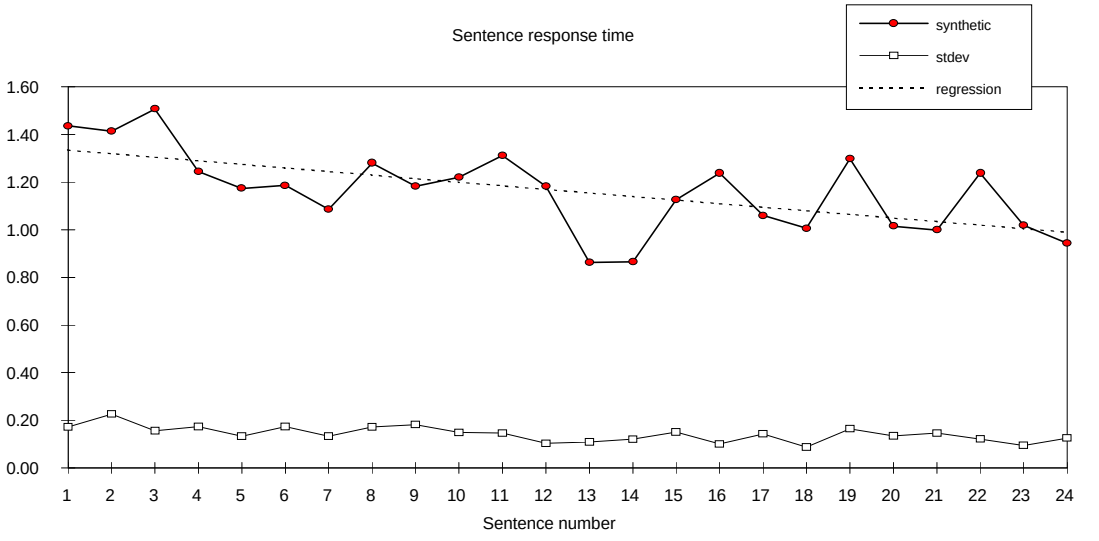

Figure 1. The mean sentence-by-sentence listening time for the practice passage. (33 subjects) and its regression line. The s.d. of the means for the individual sentences are given at the bottom of the graph.

SBSLT in the main experiment

Due to the relatively small size of the experiment, only means across all sentences in the passage are analysed for the main experiment. In Figure 2. the means for SBSLT are pooled for "easy" and "difficult" passages presented in synthetic or natural speech. Looking at the natural speech data (unfilled bars) there is a clear increase in the SBSLT from 1.0 sec to 1.5 sec going from easy to difficult. This is in accord with expectations and results in the Ralston et al. study. For the synthetic speech no significant increase is observed. Furthermore the very clear increase from natural to synthetic that was the most conspicuous result in the earlier study was not reproduced. For the easy passages this seems to be the result, but for the difficult ones the result points in the other direction. The mean SBSLT is also greater in our study, varying from 1.0 to 1.5 sec compared to about 0.6 to 0.9 sec for the earlier study. It is not clear why our subjects behaved differently. Another difference is the quality of the synthesiser. The Votrax is a commercial, low cost synthesiser with rather low quality, while we used the KTH state-of-the art text-to-speech research system. Still there is a very clear general quality difference between this system and natural speech, e.g. the segmental intelligibility, as measured according to the SAM-VCV procedure shows somewhat lower intelligibility 8.7 % errors as compared to 5.6% for natural speech (Goldstein & Till, 1992). Another factor that might affect the result is the difference in speaking rate between the synthetic and natural speech. This information is not supplied in the Ralston et al. report. In our case the natural speaker was almost 20% faster than the synthesizer, both talking at their "default speaking rate".

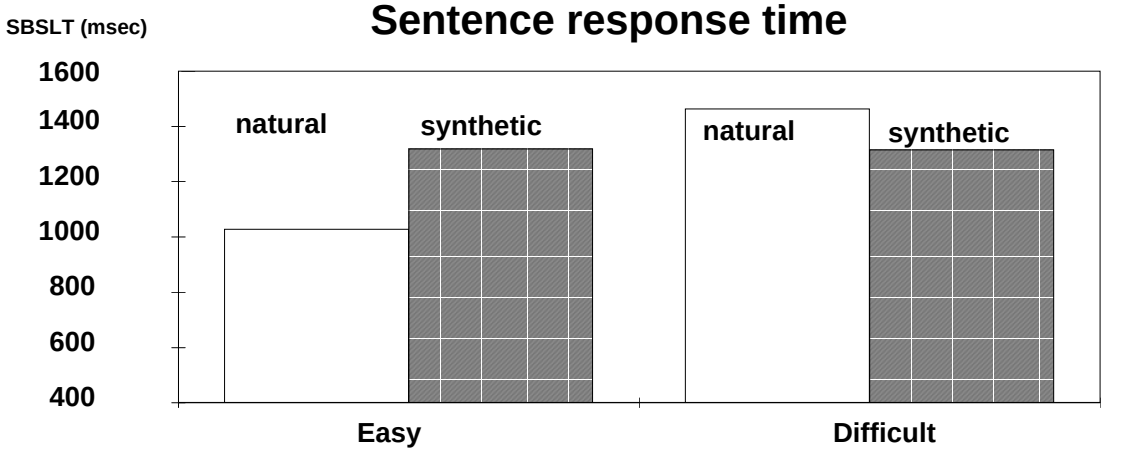

Figure 2. Mean sentence-by-sentence listening times (SBSLT) for the test passages.

Response accuracy

In Figure 3 the response accuracy in the word and proposition recognition tasks are displayed. As can be seen from the figures the results are very similar for the two tasks. There is a small (non-significant) decrease in performance from the easy to the difficult passages. Contrary to our intuition the score for the synthesiser is higher than for the natural speech. Our results for natural speech is just slightly lower (78-88%) than the earlier study (83-94%), suggesting that this part of the task is carried out in a similar fashion and that the translated test material is valid. The synthesis, in all cases, shows a superior performance on these tasks (89-95%) contrary to case in the Ralston et al. study where the synthesis performed significantly worse (72-80%). The quality of the synthesis in combination with the longer time to process the passage content, due to the slower speaking rate, seems again to account for the result.

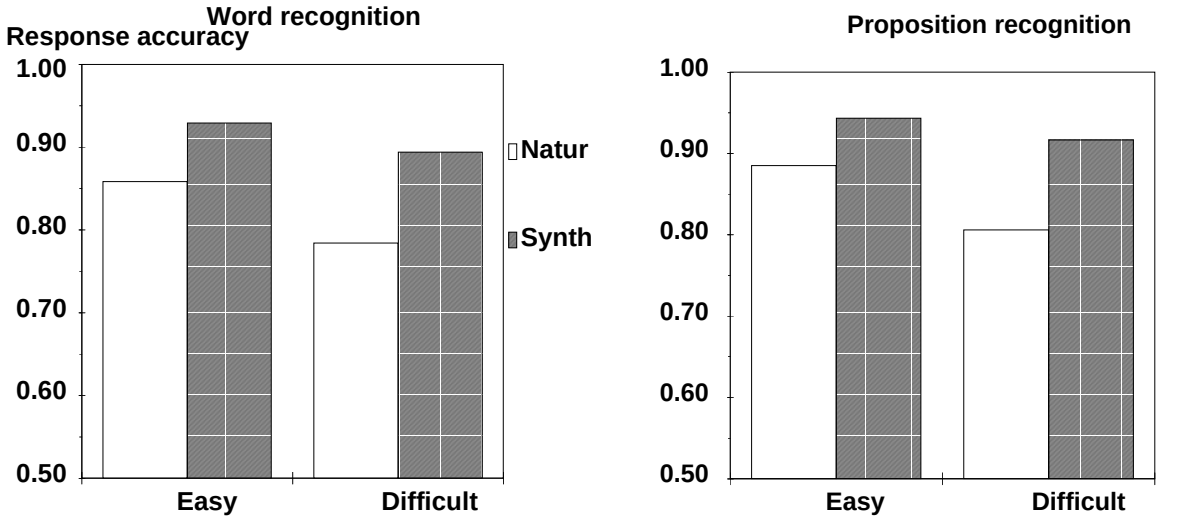

Figure 3. Mean response accuracy in the word (left) and proposition (right) recognition task.

In conclusion it seems that this test, in our current implementation, is not sensitive enough to show significant differences between high quality synthetic speech and natural speech. We might find several reasons for the discrepancy between our result and the published data by Ralston et al. Speaking tempo, speech quality, subject instructions, choice of subjects might play an important role.

ACKNOWLEDGEMENTS

We thank David Pisoni and Scott Lively for support and guidance in the implementation and adaptation of this new paradigm. This work was partially supported by grants from Swedish Telecom and the Swedish Language Technology Programme.

REFERENCES

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

Carlson, R., Granström, B., & Nord, L. (1992): "Segmental evaluation using the Esprit/SAM test procedures and monosyllabic words", *Talking machines: Theories, Models and Applications* (G.Bailly & C.Benoit, eds.)

Goldstein, M. and Till, O. (1992): Assessing segmental intelligibility of two rule-based synthesisers and Natural Speech using the ESPRIT/SAM VCV test procedures (SOAP v3.0) in Swedish and testing for differences between two correlated proportions. SAM report January 30, 1992

Logan, J.S., Greene, B.G. & Pisoni, D.B. (1989): "Segmental intelligibility of synthetic speech produced by rule", J. Acoust. Soc. Am., 86 (2), pp. 566-581

Ralston, J., Pisoni, D., Lively, S., Green, B. and Mullennix, J. (1990): "Comprehension of synthetic speech produced by rule: word monitoring and sentence-by-sentence listening times", Research on Speech Perception Progress Report, nr 16, Indiana University, pp. 119-154.