ISCA Archive
http://www.isca-speech.org/archive

The Third ESCA/COCOSDA Workshop
(ETRW) on Speech Synthesis
Jenolan Caves House, Blue Mountains,
NSW, Australia, November 26-29, 1998

Comparing the comprehensibility of different synthetic voices in a dual task experiment

Gerit P.Sonntag, Thomas Portele, Felicitas Haas

Institut für Kommunikationsforschung und Phonetik (IKP), Universität Bonn
[sonntag][portele][haas]@ikp.uni-bonn.de

ABSTRACT

We measured the comprehensibility of six German speech synthesis systems and one human voice in a dual task experiment that simulated the complexity of a real life task. PCM (Pulse Coded Modulation) and simulated GSM (Global System for Mobile communications) coding were compared. Both primary and secondary task showed significant differences in response times between voice types. Differences between the two coding conditions were significant for two of the seven voice types. Speaking rate was found to be an important factor for comprehension under increased difficulties.

1. INTRODUCTION

When assessing the output of speech synthesis systems, there are two major aspects to be evaluated: intelligibility and comprehensibility. Intelligibility has been measured in several straightforward testing methods on different levels. Nowadays high quality speech systems are known to have pretty high intelligibility rates under normal conditions. A dictation experiment (Rietveld et al., 1997) which compared three synthesis systems in two transmission conditions resulted in significantly smaller intelligibility ratings for the GSM condition (in comparison to a normal telephone condition, PSTN). Comprehensibility on the other hand cannot be measured as easily. In order to quantify the effort of speech comprehensibility, we suggest to take a closer look at the internal processing of synthetic speech stimuli.

Research on the perception and comprehension of synthetic speech suggests that there is a substantial delay in the processing of synthetic speech as compared to natural speech (Duffy & Pisoni, 1992). It is a well accepted assumption that human short-term memory is limited in its capacity to hold and process information (Luce et al., 1983). Several reaction time experiments have shown significant differences between synthetic and natural speech stimuli. However, many of these experiments have required relatively little processing effort of the listener. In a lexical decision task subjects had to decide whether they perceived a word or a non-word, or they had to repeat the word they perceived (both reported in Pisoni et al., 1985). Multiple choice tests required the subject to choose the appropriate answer to a question relating to a formerly presented speech stimulus (Sydeserff et al., 1991; Delogu et al., 1998). In a classification task subjects had to decide whether the stimulus was either human or synthetic (Nusbaum et al., 1995). Mackie et al. (1987) proposed to measure response latency and accuracy in forced choice reaction time tests for phoneme recognition. In order to increase the processing effort multiple task experiments have been performed. Luce et al (1983) carefully controlled the level of difficulty of the primary and secondary task in word and digit recall experiments and concluded that one reason for the poorer performance of synthetic speech is located in relatively early stages of pattern recognition required for word identification. But also prosodic factors like speech rate and pitch contour influence synthetic speech intelligibility and response latencies (Slowiaczek & Nusbaum, 1985; Marics & Williges, 1988). Other multiple task experiments include target word monitoring or click detection (Ralston et al., 1991; Delogu et al. 1998).

In real life applications, however, it is more likely that input which is processed by other than auditory senses has to compete with the auditory processing. Multiple task experiments with a secondary task involving different input channels have been a mouse tracking task (Boogart & Silverman, 1992) and a brick sorting task (Lazarus-Mainka & Reck, 1986) during speech presentation. In the former experiment two synthesis systems were compared with one human voice. No significant differences in the performance between voice types for either task could be measured. Comprehension of the speech stimuli was measured by questions which had to be answered by “yes/no/can’t tell”, repectively. The authors have indicated that in some cases the answer had been correct even though there had been an obvious misunderstanding. Lazarus-Mainka & Reck (1986) compared spoken and shouted natural speech. Comprehension was measured as the number of sentences correctly repeated and proved to be better for spoken than for shouted speech.

2. MOTIVATION

We wanted to compare accuracy and response latency for both a primary auditory-verbal task and a secondary visual-motor task in a more real life situation. A possible real life scenario could be a car driver who gets spoken advice on driving directions. These driving directions can nowadays be synthetic speech that is controlled automatically and sent from a remote location via radio contact. The driver then has to react to the current traffic conditions, listen to the directions and act according to both tasks simultaneously. The processing effort of such a situation is rather high and performance of both tasks may depend on the quality of the speech giving the directions.

We decided to measure speech comprehension as the reaction time of the answer given to a simple calculation task. This requires much more processing effort than a mere repetition of a speech stimulus, and the response latency is not influenced by the number of multiple choice answers or by the understanding of a question to be answered. The secondary task was to strike -as quickly as possible- one out of four coloured keys, which matched the randomly displayed colour on the monitor. The advantages of the proposed experimental set-up are obvious: every answer can unambiguously be assigned a true or false value. The stimuli of each system all differ in their meaning, but have a similar form (i.e. same words, same syntactic structure). And also the comprehension effort is not influenced by world knowledge, i.e. the individual calculation steps require the same restricted background knowledge for all the subjects.

In order to find out whether high end speech synthesis systems still require more processing capacity than human speech and whether there are measurable differences for the transmission quality of digital telephone networks we looked for differences in accuracy and response latency for the different voice types and for the two transmission conditions.

3. EXPERIMENTAL SET-UP

3.1 Stimuli and subjects

The stimuli comprised utterances from six German synthesis systems and one human voice. In order to collect the stimuli independently from the system developers they were either downloaded from interactive web sites or generated by freely available demo versions. They were calibrated to have the same mean intensity and were digitally stored (16bit, 16kHz; one system with 12kHz). One synthesis system (system ‘f‘) was of obviously poorer quality than the others. The default settings for voice type, speed, pitch etc. were not changed. So we dealt with one female and five male synthetic voices. The human voice that had been recorded in an anechoic chamber was female.

For the second experiment all utterances had undergone a simulation of a GSM coded transmission (the simulation was done by DeTeMobil Deutsche Telekom MobilNet GmbH Bonn). GSM (Global System for Mobile communications as described in ETSI, 1997) is a widespread standard for the digital cellular communication network. As the human utterances had been recorded in an anechoic chamber and as naturalness had been the main objective, the speaker had made no attempt of an especially clear pronunciation. As a consequence the human utterances were less intelligible than the synthetic utterances after the signal manipulation. Therefore we did a new recording with a more careful pronunciation. This was achieved by simulating a disturbed telephone communication channel during the recording. As a result the human utterances of the second experiment were pronounced more slowly. Mean speech rates of the synthetic versions were between 3.9 and 5.4 syllables/sec and the human utterances were 4.3 syll/sec on the average for the first experiment and 3.3 syll/sec for the second (see Figure 5).

Each utterance was a simple calculation task. We tried to put the operand always at the end of the sentence, e.g. “To the current value please add two.” This was necessary to avoid answers before the end of the request. A set of calculations comprised an initiating phrase that requested the repetition of the starting value and twelve calculation steps in a row. Each set comprised six additions and six subtractions of the operands 1 to 12. The correct results of each calculation step were all positive values and below 50. The individual calculation steps within a set were all worded differently (see Table 1). The eight sets had different successions of calculation steps and different starting values. For each of the seven different systems (for simplification we will also call the human voice ‘system’) two different sets of calculations (version a and version b) were stored.

In each experiment 21 subjects took part. 12 women and 9 men participated in the first experiment, 8 women and 13 men in the second. They were university students of different faculties, none of them was familiar with synthetic speech, and they were paid for their participation. For both experiments subjects were divided into two groups, they listened to versions a and b, respectively. Presentation order of the different systems was balanced (with 21 subjects each of the 7 systems occured three times at each position).

1. Bitte wiederholen Sie als Ausgangswert die Zahl 15.
2. Zuerst rechnen Sie plus 3.
3. Ziehen Sie bitte 8 ab.
4. Rechnen Sie eine Addition von 10.
5. Addieren Sie 4.
6. Rechnen Sie minus 9.
7. Jetzt plus 5.
8. Nun den errechneten Wert weniger 1.
9. Fügen Sie 12 hinzu.
10. Subtrahieren Sie vom Zwischenergebnis 7.
11. Geben Sie zum aktuellen Resultat 11 dazu.
12. Noch das momentane Ergebnis minus 2.
13. Rechnen Sie eine Subtraktion von 6.

Table 1: Example of one set of calculations.

3.2 The different tasks

The subjects were asked to do two things at the same time: respond to a colour displayed on a monitor by a keystroke and respond to a calculation request via headset by answering it. These two tasks which may seem simple at first sight are indeed rather difficult. A couple of informal preliminary test runs with calculation results over 100, including multiplications and divisions, were obviously too complicated. We then tried to simplify the calculations by excluding multiplications and divisions and by staying within the range of 0-50. We are nevertheless dealing with a high workload for the short term memory. Figure 1 schematizes the subjects´ mental workload: there is not only the processing of the two input stimuli, the processing of the appropriate responses, but also the calculated value has to be kept in memory until the next calculation step. There was no alignment in time between the two tasks, the next colour display was triggered by the subject’s keystroke, and the test administrator who was seated outside the chamber started the playback of the next request as soon as the answer was given (whether it was correct or wrong).

3.3 Procedure

The subjects were told that this was a reaction time experiment and that it did not matter, whether their calculations were correct or not. They were instructed that, in case they lost track of the current value, they were to keep going on from a random value. These cases were counted as just one mistake afterwards. The subjects were seated in an anechoic chamber to ensure high quality recordings of their spoken answers. In the chamber was a computer for the colour task with a prepared keyboard: four adjacent keys were coloured in blue, green, yellow and red. At first the subjects were asked to do a practice run of the colour task; they were free to use one hand or both. A second practice run included both tasks. The actual test consisted of seven calculation sets of the seven different systems; it lasted around 15 minutes altogether. After each set of calculations the subjects

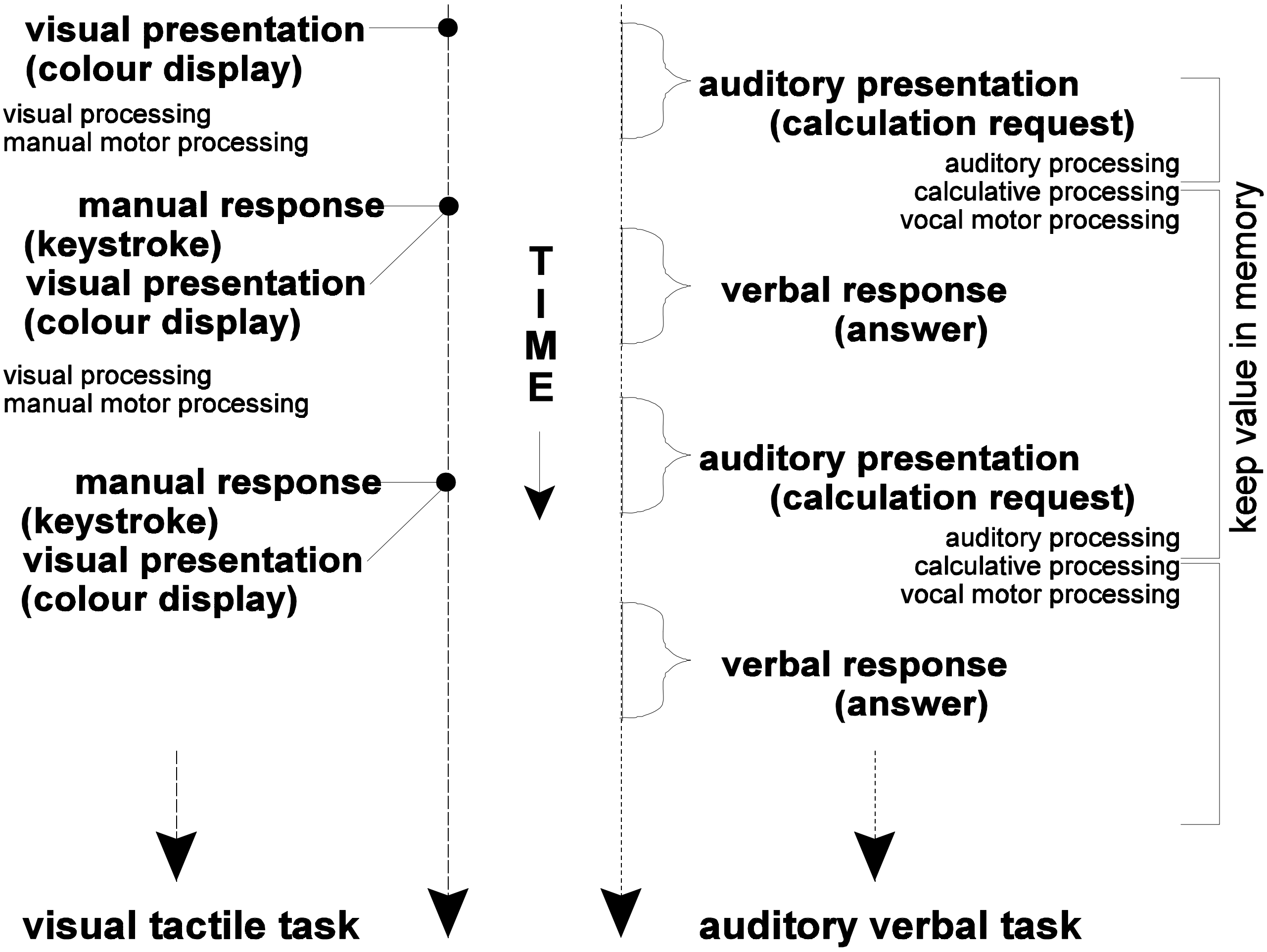

Figure 1: Temporal scheme of the different tasks.

were given feedback information about the accuracy of each set‘s final result. These feedback utterances were prerecorded human utterances and motivated the subjects to calculate correctly. The test administrator chose the appropriate feedback information. The subject did not hear anything apart from the prerecorded utterances and there was no eye contact with the test administrator. If necessary the subject could ask for a repetition of an individual request.

3.4 Pretest

A pretest was carried out to ensure that differences between systems will not be due to varying difficulty of the sets of calculations. To analyse whether all sets were of equal difficulty they were recorded with a human voice and presented to five subjects. The subjects did a trial run of the calculation task and of both tasks in combination. There was a time span of several days in between the experiments and the subjects reported that they did not remember the calculations of the first run while performing the second run. For each subject the presentation order of the eight different sets was changed at random.

Figure 1 also shows that the visual-tactile task was recorded as several *points in time*, whereas the auditory verbal task consists of *time intervals*: the calculation request, the interval of the internal processing, the spoken answer and the interval until the next request. How many keystrokes fell into each interval was entirely up to the subject‘s performance.

The recorded “dialogues” were automatically segmented into utterances and pauses and the labels were manually corrected. The results of the colour task were aligned in time with the recordings, so that each keystroke can be allocated a certain point within the “dialogue”. The reaction time of the calculation task is defined as the pause between the calculation request and the following answer.

The reaction time of the colour task is defined as the time between two keystrokes. To exclude reaction times of keystroke intervals that were well outside the time of the internal auditive verbal processing we only considered keystroke intervals that fell within either the presentation of the request or the following pause. We also excluded all keystroke reaction times that exceeded two “dialogue” intervals.

In the pretest 2.5% of the calculations were wrong and in 2.9% of all cases subjects struck the wrong colour key. The mean answer reaction time without the colour task was 1074ms (stdev:1040ms) and 1585ms (stdev:1422ms) with the colour task. The mean keystroke reaction time was 1281ms (stdev: 705ms). For all these reaction times the differences between subjects were significant (ANOVA, $p<0.001$). In order to analyse the effect of the different sets without the effect of the subjects, all reaction times were converted into *z-values*. After the transformation each subject had a mean of 0 and a standard deviation of 1. An analysis of variance was carried out for the effect of the calculation set on reaction time and accuracy of both tasks. There was no significant effect on the answer reaction time (F(7,512)=0.989), nor on calculation accuracy (F(7,408)=0.765), nor for colour accuracy (F(7,741)=1.939) (all computed with *z-values*). The effect on keystroke reaction time (F(7,741)=2.087, p=0.043) can be interpreted to be significant, but a pairwise comparison (Scheffé and Tukey) between the individual sets showed no significant differences at the 0.05 level. As a consequence, we assume all sets of calculations to be of equal difficulty.

4. RESULTS

Reaction times were measured as described in section 3.4. 90% of all answer reaction times were below 2500ms and 90% of all keystroke reaction times were below 1700 ms (see Fig. 2).

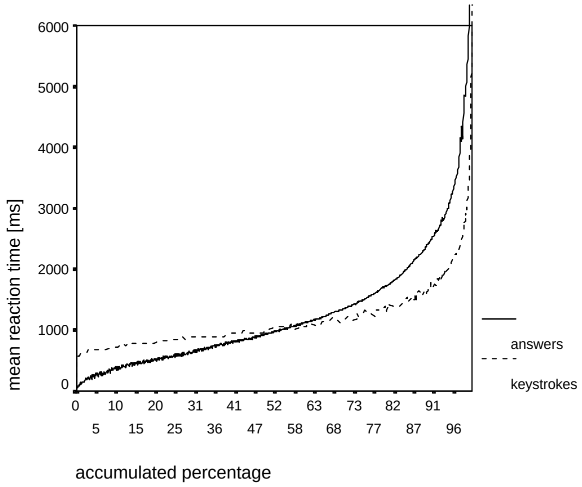

Figure 2: Distribution of measured reaction times of both experiments.

4.1 The calculation task

7% (exp.1; exp.2: 6.2%) of all calculation steps were wrong. Most of the wrong calculations were due to system ‘f‘, which accounted for 30% (exp.1) and 42% (exp.2) of the wrong answers. The other systems did not significantly affect the accuracy of calculation. In both experiments reaction time of wrong calculations was significantly (p<0.01) longer than the time needed for correct answers.

An analysis of variance indicated that the effect of subject was significant for the reaction times in both experiments (exp.1: F(20,1863)=6.047, p<0.001; exp.2: F(20,1890)=11.148, p<0.001). Female subjects responded significantly more slowly than male subjects and in exp.1 they even made significantly more mistakes. Subjects’ mean response times varied between min=800ms and max=1852ms (exp.1, N=1884), min=547ms and max=2422ms (exp.2, N=1911). In order to exclude the effect of inter-subject differences, the *z-values* of the subjects‘ reaction times were computed for both experiments. A global comparison of the *z-values* of the two experiments indicated that there were no significant differences between the two conditions. Looking at each system individually, we found a significant increase in reaction time for system ‘c‘ (F(1,531)=7.492, p<0.01) for the GSM condition. Reaction times for the human voice were significantly shorter for the GSM condition (F(1,544)=5.511, p<0.05) (see Figure 3). An analysis of variance indicated a significant effect of system on the reaction times (exp.1: F(6,1877)=4.715, p<0.001; exp.2: F(6,1904)=10.480, p<0.001). However, a Scheffé pairwise comparison at the 0.05 level revealed that in the first experiment only the difference between the human voice and system ‘f‘ was significant. In experiment two the human voice differed significantly from all the other voices, but there were no significant differences in between the synthetic voices. If we pool the results of the two experiments, then the differences between the human and all synthetic voices become significant (p<0.05), just as the differences between system ‘f‘ and all other systems apart from system ‘c‘.

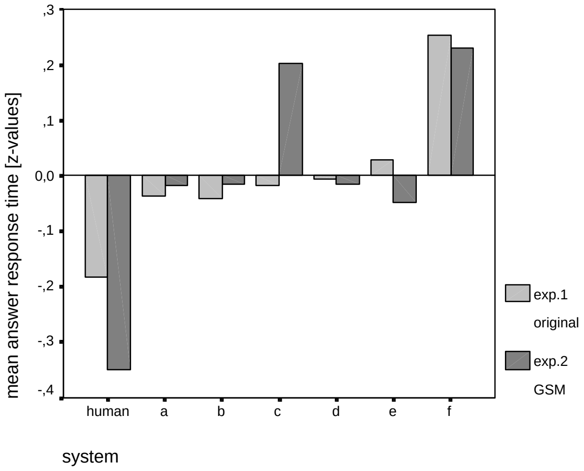

Figure 3: Mean reaction times of the calculation task across the seven systems; indicated are *z-values*.

4.3. The colour task

Subjects struck the wrong colour key in only 1.7% (exp.1) and 6.4% (exp.2) of all cases. Again most of the mistakes were due to system ‘f‘ (33% exp.1 and 42% exp.2), and the other systems did not significantly affect accuracy. There were no significant influences of the colour presented on keystroke reaction time, on answer reaction time, nor on the correctness of calculation.

Subjects’ mean response times varied between min=981ms and max=1328ms (exp.1, N=2550), min=909ms and max=1426ms (exp.2, N=2076). Again we computed the *z-values* to take into account the significant differences between subjects (exp.1: F(20,2529)=5.919, p<0.001; exp.2: F(19[1],2056)=7.719, p<0.001). Comparing the two conditions we found no significant differences, not even when looking at the different systems individually. So we pooled the results of the two experiments and then checked for differences between the voices. The significant effect of system on reaction time (F(6,4619)=6.559, p<0.001) was further analysed by a Scheffé pairwise comparison. Only the differences between system‘f‘/system‘a‘ and system‘f‘/human voice were indeed significant (p<0.05)(see Figure 4).

[1] One subject was so slow in his keystroke reactions that he was excluded from further analysis by the filtering for relevant keystrokes (as defined in 3.4).

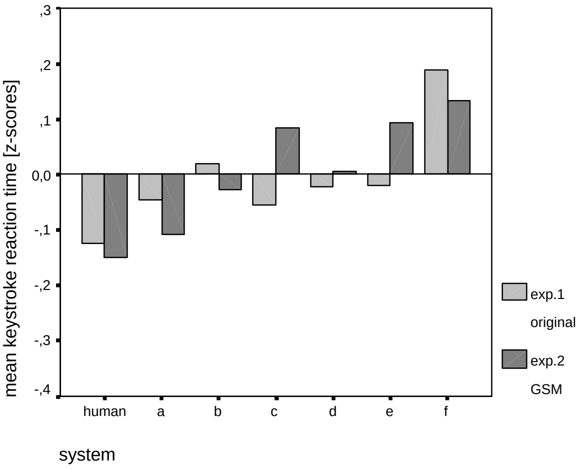

Figure 4: Mean reaction times of the colour task across the seven systems; indicated are *z-values.*

5. DISCUSSION

We could show differences in processing time between different types of voices for both primary and secondary task. However, the differences between the top quality synthesis systems and the human voice were not always significant. The pooled results of both transmission conditions show that answer reaction times for the human voice were significantly shorter than those for the synthetic voices. For the secondary task the human voice differed significantly only from the low end synthetic voice. Thus we can distinguish between the human voice, five top quality systems, and one low end system. The only observable difference between the top quality systems was the poor performance of system ‘c‘, which was due to the GSM condition.

There was no general shift in performance for the two different transmission conditions. The only noticable differences were the increase in performance of the human voice and the decrease of system ‘c‘ for the GSM condition. Differences of the human voice can be explained by the use of different stimuli. As has been shown in the literature (van Summers et al., 1988), speakers attempt to maintain a constant level of intelligibility in the face of degradation of the message by environmental noise. We tried to simulate this effect, known as the Lombard reflex, in order to have a realistic standard for the comparison of the synthetic voices. However we did not expect performance for the human voice to improve for the more difficult GSM condition. The improvement of the comprehensibility between the first and second recording seemed to be bigger than the degradation due to the coding condition. One of the human strategies to improve intellibility is known to be a decrease in speaking rate. The second recordings with the simulated disturbed communication channel were indeed significantly slower (p=0.001) than the first recordings in silence (see Figure 5). The effect of speaking rate may also explain the poor performance of system ‘c‘, as it had a significantly higher speaking rate (mean: 5.4 syll/sec) than the other systems (mean: 4.3 syll/sec). As this difference seemed not to affect performance under the PCM condition, we conclude that the effect of speaking rate becomes more important with increased comprehension difficulty. The important influence of speaking rate on the intelligibility of synthetic speech in adverse conditions has also been observed by Köster & Mersdorf (1998).

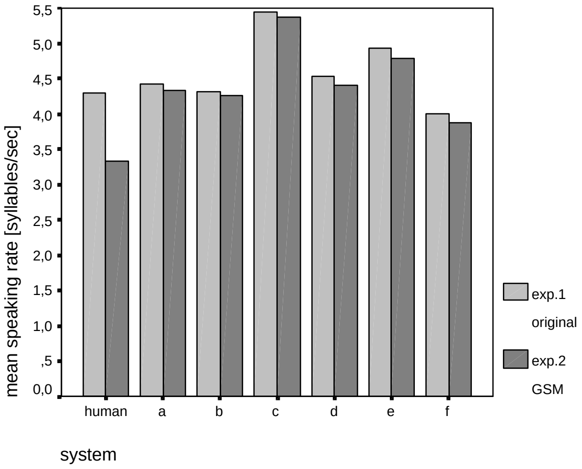

Figure 5: Speaking rate of the different systems, only the human voice stimuli were different for the two experiments. The other differences between experiments are due to the manipulation.

The measure of accuracy did not differentiate between the top quality systems due to a ceiling effect. Only the low end system could be significantly distinguished from the human voice. However, we think that the demands placed on the subjects were so high, that increasing the complexity of the task will not help to better differentiate between systems.

After all we have shown how close nowadays high end synthesis systems come to the human voice, not only as far as intelligibility scores are concerned but also when it comes to comprehensibility. The rather elaborate experiment described in this paper showed differences in comprehensibility, but failed to differentiate significantly between top quality systems. Yet, we still believe that experiments which simulate more real life tasks will be useful even for comparative evaluation.

For further research we would like to suggest two main aims. On the one hand, it would be interesting what adaptation strategies human speakers employ under difficult conditions and how they can be copied to improve synthesis performance (cf. Köster & Mersdorf, 1998). On the other hand, as comprehension differences between human and synthetic voices seem to decrease, other dimensions of speech quality should come more into focus. In order to differentiate even between top quality synthetic voices, dimensions like pleasantness and acceptance should be of primary importance both for short term applications (e.g. verbal directions) and for long term applications (e.g. newspaper reading).

6. ACKNOWLEDGEMENTS

We would like to thank DeTeMobil Deutsche Telekom MobilNet GmbH for the processing of the stimuli, Karlheinz Stöber for the programming of the automatic pause detector and of the results’ processing and, most of all, our subjects for their participation and patience.

7. REFERENCES

Boogart,T.; Silverman,K. (1992) “Evaluating the overall comprehensibility of speech synthesizers”, in: Proceedings of the International Conference on Spoken Language Processing (ICSLP), Alberta, Canada, vol.2, pp.1207-1210

Delogu,C.; Conte,S.; Sementina,C. (1998) “Cognitive factors in the evaluation of synthetic speech”, in: Speech Communication 24, pp.153-168

Duffy,S.A.; Pisoni,D.B. (1992) “Comprehension of synthetic speech produced by rule: a review and theoretical interpretation”, in: Language and Speech 35, pp.351-389

ETSI (European Telecommunications Standards Institute) (1997) “Digital cellular telecommunications system”, GSM 06 series

Köster,S.; Mersdorf,J. (1998) “Verstehbarkeit von Sprachsynthese gehört über Telefon in lärmbehafteter Umgebung ”, Tagungsband der 9. Konferenz Elektronische Sprachsignalverarbeitung/ITG Fachbericht Sprachkommunikation 152, Dresden, pp.81-84

Lazarus-Mainka,G.; Reck,S. (1986) “Sprachverständlichkeit als Funktion der Prosodie” in: Zeitschrift für Psychologie 194(2), pp.191-204

Luce,P.A.; Feustel,T.C.; Pisoni,D.B. (1983) “Capacity demands in short-term memory for synthetic and natural speech”, in: Human Factors 25(1), pp.17-32

Mackie,K.; Dermody,P.; Katsch,R. (1987) “Assessment of evaluation measures for processed speech”, in: Speech Communication 6, pp.309-316

Nusbaum,H.C.; Francis,A.L.; Henly,A.S. (1995) “Measuring the Naturalness of Synthetic Speech”, in: International Journal of Speech Technology 1, pp.7-19

Pisoni,D.B.; Nusbaum,H.C.; Greene,B.G. (1985) “ Perception of Synthetic Speech Generated by Rule ”, in: Proceedings of the ICEE 73(11), pp.1665-1685

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

Rietveld,T.; Kerkhoff,J.; Emons,M.J.W.M.; Meijer,E.F.; Sanderman,A.A.; Sluijter,A.M.C (1997) “Evaluation of speech synthesis systems for Dutch in telecommunication applications in GSM and PSTN networks” in: Proceedings of Eurospeech, vol.2, pp.577-580, Rhodes, Greece

van Summers,W.; Pisoni,D.B.; Bernacki,R.H.; Pedlow, R.I.; Stokes,M.A. (1988) “ Effects of Noise on Speech Production: acoustic and perceptual analysis ”, in: Journal of the Acoustical Society of America 84(3), pp.917-928

Slowiaczek,L.M.; Nusbaum,H.C. (1985) “Effects of Speech Rate and Pitch Contour on the Perception of Synthetic Speech”, in: Human Factors 27(6), pp.701-712

Sydeserff,H.A.; Caley,R.J.; Isard,S.D.; Jack,M.A.; Monaghan,A.I.C. (1992) “Evaluation of Speech Synthesis Techniques in a Comprehensive Task”, in: Speech Communication 11, pp.189-194