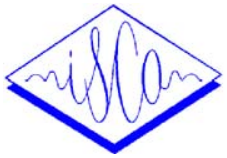

EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

# EVALUATION OF SPEECH SYNTHESIS SYSTEMS FOR DUTCH IN TELECOMMUNICATION APPLICATIONS IN GSM AND PSTN NETWORKS

*T. Rietveld (1), J. Kerkhoff (1), M.J.W.M. Emons (2), E.J. Meijer (2), A.A. Sanderman (2) A.M.C. Sluijter (2).*

(1) University of Nijmegen, the Netherlands, Erasmusplein 1, 6525 HT Nijmegen, The Netherlands,
Tel. +31 24 3612905, E-mail: a.rietveld@let.kun.nl
(2) KPN Research, Leidschendam, the Netherlands



## ABSTRACT

In this contribution the subjective evaluation of three Text-To-Speech systems (two diphone and one allophone system) is reported in three transmission conditions: standard telephone (PSTN) and GSM. The three TTS-systems realised three different texts: Travel information, Stock Exchange Reports and E-mail messages. The subjects had to carry out three tasks: a) to give preference judgements on the three TTS-systems and b) to rate the readings on 16 five-point scales. The rankorder on the scale of general quality was: Public Transport > Stock Exchange > E-mail reading, in both transmission conditions. The GSM-transmission tends to decrease the perceptual scores on a number of subjective scales, In the transliteration task significantly more errors were made in the GSM-condition than in the PSTN-condition. In both conditions less errors were made with the diphone TTS-systems than with the allophone system.

## 1. INTRODUCTION

Nowadays, the number of voice response services is increasing rapidly. Customers tend to get used to these services more and more. In the beginning a lot of resistance against DTMF services was observed, but nowadays customers take advantage of the added value provided by these services and their 24-hour availability. Since the customer's attitude towards these services has undergone a positive change, companies wish to increase the amount of information supplied through these systems. Customer service improves and costs reduce.

In current voice response services in the Netherlands only speech concatenation techniques are used. These techniques have the disadvantage of a rather low level of flexibility of the service because of their dependence on pre-recorded speech items. Moreover, for some services it is simply impossible to pre-record all items because of the enormous diversity of the speech output. Therefore time has come to introduce speech synthesis in order to make any possible kind of information available and to allow us to build much more flexible systems than hitherto. Future telecommunications services such as reverse directory, messaging services can't do without text-to-speech synthesis.

### 1.1 Research questions

At this moment there are several Text-To-Speech systems for Dutch available. In the investigation reported here we evaluated one allophone-based speech synthesiser and two diphone-based systems, realising three types of texts:
a) travel information, b) stock exchange reports and c) e-mail, in two conditions: GSM and PSTN.

The experiments reported here evaluate the quality and acceptability of the available systems in order to determine if they are yet suitable for use in specific applications. In the project we are reporting here we focus on two different issues:

a. It was assessed whether for applications such as e-mail reading the quality of speech synthesis is inferior to the added value of the service.
b. It was also assessed whether the quality of speech synthesis is preserved in analog (PSTN) and digital (GSM) telephone networks. Especially in a mobile environment people do feel the need to access their e-mail, or to hear traffic news by telephone because they have less other ways to get access to this information.

### 1.2. Experimental set-up

The evaluation consisted of two parts:
I Subjective evaluation: listeners representing potential customers are asked to express their judgements on listening comfort and system preference.
II Intelligibility: the intelligibility of the systems is tested on the basis of semantically unpredictable sentences containing all Dutch vowels, consonants and clusters of two consonants.

The materials presented in parts I and II consist of the three text types mentioned in the introduction; they are both performed in simulated GSM and PSTN conditions. 44 subjects took part in part I and also 44 in part II. Twenty-two subjects participated in the experiment with PSTN-transmitted materials and 22 others in the experiment in the GSM-condition.

## 2. DESIGN

The subjects were given two tasks: preference judgements between readings and a scaling task. Another

group of subjects was given the task to transliterate a semantically unpredictable text.

1) *Preference judgements:*

Preference of reading (a) to reading (b), expressed on a seven-point scale.

2) *Subjective judgements on 16 five-point scales*

2.1 general quality
2.2 listening effort
2.3 comprehension problems
2.4 intelligibility
2.5 pronunciation
2.6 speaking rate
2.7 voice pleasantness
2.8 naturalness
2.9 liveliness
2.10 relaxed speaking style
2.11 quiet speaking style
2.12 friendliness
2.13 fluency
2.14 politeness
2.15 personality
2.16 calling without resistance

Items 2.1 - 2.6 and 2.16 are in agreement with the ITU-standards [3]. In the ITU-standards voice pleasantness is just one item, but in our test it was split up in nine items (2.7 - 2.15).

In this experiment three texts generated by three different Text-To-Speech systems were evaluated. The set of systems at issue consisted of two diphone speech synthesisers (labelled D1 and D2), and one allophone synthesis system (labelled AL). The allophone TTS-system is a sequential modular system based on the use of a Central Data Structure with a great number of aligned information streams [2]. One of the two diphone systems (D1) is a Diphone phonetics-to-speech system; it accepts enriched SAMPA-phonemic representations of text from any particular application. The enrichments concern specifications for the location of accents and prosodic boundaries. The programme concatenates the corresponding diphones (either waveform or LPC-diphones), and generates a pitch contour anf phoneme durations on the basis of specifications in files containing rule sets. This system uses the linguistic and phonological information which is available in the allophone TTS-system. The other diphone system (D2) is a Text-To-Speech system.

The Public Transport text comprised 88 words (365 letters), the E-mail text 101 words (437 letters) and the text on the Stock Exchange 87 words (502 letters).

A crucial effect in the evaluation of synthesis systems is that repeated presentation of readings results in familiarity, which, in turn, may result in higher intelligibility, and consequently in a more positive attitude towards the system at issue. That is why we counterbalanced the presentation of pairs of texts over the subjects. To that end a Greek-Latin Square was used to assign the subjects to the conditions.

The transmission conditions (PSTN and GSM) are simulated in the following way: In the PSTN-condition the synthesised speech fragments are amplified and fed to a normal telephone receiver via the IRS system which simulates the transmission characteristics of a normal telephone. In the GSM-condition the speech fragments are played via a coupling with the analog network to a GSM-receiver with car-kit (Pocketline Newton). Background noise was added which had been recorded in a car (70 ($\pm$3) dB SPL).

3) *Intelligibility test*

In this test the subjects were asked to transliterate a semantically unpredictable text of 147 words; articles like 'the' were not counted. Each of the subjects listened to just one reading, per condition 7 to 8 subjects took part in the experiment.

The subjects involved all had higher education; their ages ranged from 19 - 45 years.

## 3. RESULTS

### 3.1 Preference judgements

The preferences for the three Text-To-Speech systems involved were obtained with different subjects for different texts, both in the PSTN- and in the GSM-conditions. The subjects were asked to rate their preferences for pairs of systems on a scale ranging from -3 to +3. For the order of presentation 'system a/system b', a rating of -2 would mean a moderate preference for system a. The subjects were presented both orders.
The preferences obtained with the three texts were pooled and subsequently processed by a specific analysis of variance for paired comparison data (Scheffé 1952; program DIFANOVA by first author). This analysis results in a) omnibus statistical tests for the main effects, b) locations of the compared objects on an interval scale and c) a yardstick which can subsequently be used in a post-hoc comparison test to assess which objects differ significantly from each other on the preference scale. Using an alpha-level of .05, significant main effects were found for the TTS-systems both in the PSTN and the GSM-conditions. The scale values obtained by the three systems are depicted in figure 1.

In the PSTN condition the positions of the allophone synthesiser and one of the diphone synthesisers (D1) on the scale of preferences did not differ significantly; both dif

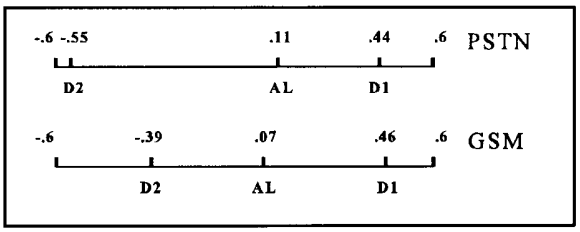


*Fig. 1 Relative positions of the three TTS-systems based on preference scaling as a function of transmission condition.*

fered significantly from diphone synthesiser D2; the size of the yardstick (= the minimal difference required for a significant difference) was .503. In the GSM-condition only one contrast was significant: the difference between diphone system D1 and diphone system D2.

### 3.2.1 The effects of TTS-system, transmission and text on scale values

The subjects were asked to scale the readings of the three texts by the three systems on 16 five-point scales. As a first step a multivariate analysis of variance was carried out on the scores of the 16 scales involved. The independent variables were TTS-system (D1, D2 and AL synthesiser), Transmission (PSTN and GSM) and Text (stock market, e-mail and public transport) and their interactions. Only two factors turned out to be significant at the 5%-level:
*Transmission*: Pillais F(16,198) = 1.82, p = 0.030 and *Text*: Pillais F(32,398) = 4.73, p = 0.000;
no interactions were significant. The GSM-condition generally yields lower scores on the subjective scales than the PSTN-condition.

The univariate F-tests associated with these two factors (the significance level for these tests was set at a lower level: 1% in order to reduce the chance of a type I error) showed that the factor Text was significant on 11 scales, whereas the factor Transmission only achieved significance on two factors: *Rate of Speech* and *'Listening effort'*. The overall-rankorder of judgements (from negative to positive) for the three text types involved was *E-mail*, *Stock-Exchange* and *Public Transport Information*. The positions taken on the scale *General Quality* (Table 1) is illustrative for the positions on the other scales:

*Table 1 Mean values on scale 'General Quality'*

| | E-mail reading | Stock-exchange | Public Transport |
|---|---|---|---|
| PSTN | 1.95 | 2.51 | 3.29 |
| GSM | 1.84 | 2.58 | 3.13 |

*Table 2 Mean scale values on 16 scales, pooled over three texts, obtained by three Text-To-Speech systems (D1 and D2 are both diphone systems, AL is an allophone synthesiser) in two transmission conditions: PSTN and GSM.*

| | PSTN | | | GSM | | |
|---|---|---|---|---|---|---|
| SCALE | D1 | D2 | AL | D1 | D2 | AL |
| General quality | 2.73 | 2.42 | 2.70 | 2.72 | 2.34 | 2.49 |
| Listening effort | 3.15 | 3.00 | 3.00 | 3.20 | 2.73 | 2.87 |
| Comprehension problems | 3.65 | 3.39 | 3.45 | 3.44 | 3.14 | 3.13 |
| Intelligibility | 2.91 | 2.74 | 2.75 | 2.91 | 2.42 | 2.65 |
| Pronunciation | 3.09 | 3.00 | 3.18 | 3.09 | 2.84 | 2.87 |
| Speaking rate | 3.18 | 3.09 | 3.16 | 2.98 | 2.81 | 2.90 |
| Voice pleasantness | 2.91 | 2.26 | 2.86 | 2.77 | 2.41 | 2.49 |
| Naturalness | 2.42 | 2.35 | 2.32 | 2.30 | 2.07 | 2.42 |
| Liveliness | 2.70 | 2.43 | 2.66 | 2.63 | 2.34 | 2.78 |
| Relaxed speaking style | 3.00 | 2.91 | 3.14 | 2.63 | 2.34 | 2.78 |
| Quiet speaking style | 3.09 | 3.04 | 2.86 | 3.55 | 3.25 | 3.84 |
| Friendliness | 3.60 | 3.22 | 3.32 | 3.58 | 3.41 | 3.18 |
| Fluency | 2.24 | 1.91 | 2.23 | 2.28 | 1.80 | 2.16 |
| Politeness | 3.70 | 3.22 | 3.61 | 3.63 | 3.63 | 3.49 |
| Personal | 2.58 | 2.48 | 2.55 | 2.67 | 2.32 | 2.73 |
| Calling without resistance | 2.91 | 2.57 | 2.84 | 3.07 | 2.41 | 2.70 |

### 3.2.2 Which variables affect perceived quality most?

In order to assess to which extent scores on the subjective scales can predict the perceived quality of the readings by the three texts, a multiple regression analysis was carried out on the scores, with the variable *general quality* as criterion, and the other scales as predictors. The method used in this analysis was 'stepwise', and the entrance criterion was set at .05.
Four predictors turned out to explain significantly the observed variation in the scores on the scale *general quality*.

*Table 3 Statistics associated with the four significant predictors for perceived quality of readings by three Text-To-Speech systems.*

| Predictor | β | T-statistic | Cumul. $R^2$ | p-value |
|---|---|---|---|---|
| Comprehension problems | .47 | 7.91 | .754 | .000 |
| Relaxed speaking style | .28 | 5.75 | .789 | .000 |
| Pleasantness | .12 | 3.01 | .806 | .003 |
| Personal | .13 | 2.37 | .811 | .019 |

Apparently the predictor 'comprehension problems' contributes most to the perceived quality; the contribution of the other predictors is almost negligible, although significant. One has to keep in mind, though, that many variables were highly correlated.

### 3.3 Intelligibility test

In table 4 we display the percentages of correct (+) and incorrect (-) transliterations of the 147 words in the text.

*Table 4 Percentages correct (+) and incorrect (-) transliterations of words by 7 to 8 listeners in each of six conditions (2 transmission conditions x 3 TTS-systems).*

| | PSTN | | | GSM | | |
|---|---|---|---|---|---|---|
| | D1 | D2 | AL | D1 | D2 | AL |
| + | 68.2 | 72.4 | 66.4 | 55.1 | 59.2 | 51.7 |
| - | 31.8 | 27.6 | 33.6 | 44.9 | 40.8 | 48.3 |

A loglinear analysis carried out on the frequencies showed that both main effects were significant at the 5%-level; no significant interaction was found. A subsequent contrast analysis revealed that the intelligibility of the two diphone TTS-systems differed significantly from that of the allophone system. The two diphone systems also differed significantly. Clearly the GSM condition decreases the intelligibility of the synthesised text, and did so for all systems involved. The two diphone systems outperformed the allophone TTS-system.

## 4. DISCUSSION AND CONCLUSION

The results obtained in this series of evaluation experiments can be summarised as follows:
The GSM-condition tends to decrease the perceptual scores on a number of subjective scales, although the test used here, in which no repeated measurement design could be used, was not a very powerful. The difference was confirmed by the results of the test in which intelligibility was measured in a transliteration task: the number of errors was significantly lower in the PSTN transmission than in the GSM transmission.

The extent to which a reading is perceived as posing comprehension problems predicts to a very large extent the perceived quality of a reading; the contribution of other predictors is negligible.

The perceived quality of a reading, as measured on the basis of a large number of subjective scales, depends on the specific text at issue. Simpler texts, without much syntactical and prosodic complexity, tend to be judged more positive than texts whose complexity may lay beyond the processing capacities of the Text-To-Speech systems at issue. In our investigation the readings of Public Transport Information only reached a point of satisfactory quality.

The scores obtained on the subjective scales by the three TTS-systems differed, but not significantly. However, there was tendency for D1 to obtain higher scores than the Allophone system, which, in turn, performed somewhat better than D2.

In the second part of the project, not reported here, we investigate the possibility to improve the quality of the three systems by developing application-specific modules. For instance, numbers in Stock Exchange reports require a different reading style than numbers in soccer reports. Although the quality of the systems might not be sufficient to generate fully free text, it might be possible to add text-specific parameters so that the quality is acceptable for certain applications. In this part the segmental and suprasegmental aspects of three different text types: e-mail, travel information and financial news are studied. In a next step the characteristic features of these texts are built in a pre-processing module to enrich the texts before reading.

## 5. REFERENCES

[1] Jongenburger, W. & van Bezooijen, R. (1992) Evaluatie van ELK: attitudes van de gebruikers, verstaanbaarheid en acceptabiliteit van de spraaksynthese. ASSP-Report.
[2] Kerkhoff, J. & Rietveld, T. (1995) The generation of prosody in the Nijmegen Rule Oriented speech synthesis system. In: Proceedings Eurospeech, Madrid, Vol. 3, 1831-1834.
[3] ITU-T Recommendation P. 85. A method for subjective performance assessment of the quality of speech voice output devices