COMPARATIVE EVALUATION OF SIX GERMAN TTS SYSTEMS

Gerit P.Sonntag, Thomas Portele, Felicitas Haas, Joachim Köhler *

Institut für Kommunikationsforschung und Phonetik (IKP)
Universität Bonn, Bonn, Germany
{sonntag, portele, haas}@ikp.uni-bonn.de
* Siemens, ICP CD TI 24, Bocholt, Germany

ABSTRACT

An application-specific perceptual evaluation was carried out in order to compare six high-quality German text-to-speech systems. Subjects judged the systems' reading of an email message and a newspaper article according to four application-specific questions and six voice quality attributes. The results indicate significant differences between the systems. Possible applications of the systems were judged rather unfavourably. The main reasons for this proved to be the synthetic prosody and voice quality. Errors concerning text conversion were less important.

1. INTRODUCTION

Today the quality of high-end text-to-speech systems cannot be adequately measured in intelligibility rates alone. Most systems achieve intelligibility scores that are close to human speech [1]. But more distinguish-able measures are strongly necessary, especially for the end user who wants to know which system will best fulfill his application specific needs. For the developer it is important to know which parts of his system need further improvement (e.g. text conversion, prosody and/or voice quality). This investigation was carried out to assess whether six leading TTS systems for German can be used in specific contexts like email reading or news announcements via telephone, and whether differences in the quality of the systems can be measured. Contrary to a functional evaluation with similar systems [3] only subjective impressions were collected because acceptance by potential customers and success in the market depend on the customers willingness to use such services.

2. STIMULI GENERATION

An email message (including email header) of 203 words and a newspaper article of 258 words were synthesized by three PSOLA-based TTS systems P1, P2, P3 (two of which are diphone based, one has a mixed inventory structure), two LPC-based systems L1 and L2 and one formant synthesizer F1. All but one system's stimuli were generated independently from the system developers by assessing interactive websites or using freely available demo versions. As only male voices were to be compared, the default settings of one system had to be changed. All stimuli were digitally stored (8kHz / 16bit). A human version of both texts was recorded in order to make the subjects familiar with the text first before judging the different synthetic voices.

Two of the authors counted the text conversion errors of each system, taking the human version as a reference. Most of the errors (table 1) were due to numbers (date/time) in the text or to proper nouns that were pronounced inadequately.

system	email text	newspaper
P1	2	10
L1	16	12
P2	8	16
P3	6	12
F1	8	10
L2	10	19

Table 1: Number of text conversion errors for each system. Systems are sorted according to their position in the category estimation.

3. TEST PROCEDURE

Category estimation (also called mean opinion score - MOS) has proven a reliable evaluation method for opinion tests. Usually a 5-point scale with verbal categories is being used [4]. We decided to employ an even-numbered 6-point scale in order to prevent the subjects from giving 'neutral' judgements in the middle of the scale. The scales were constructed
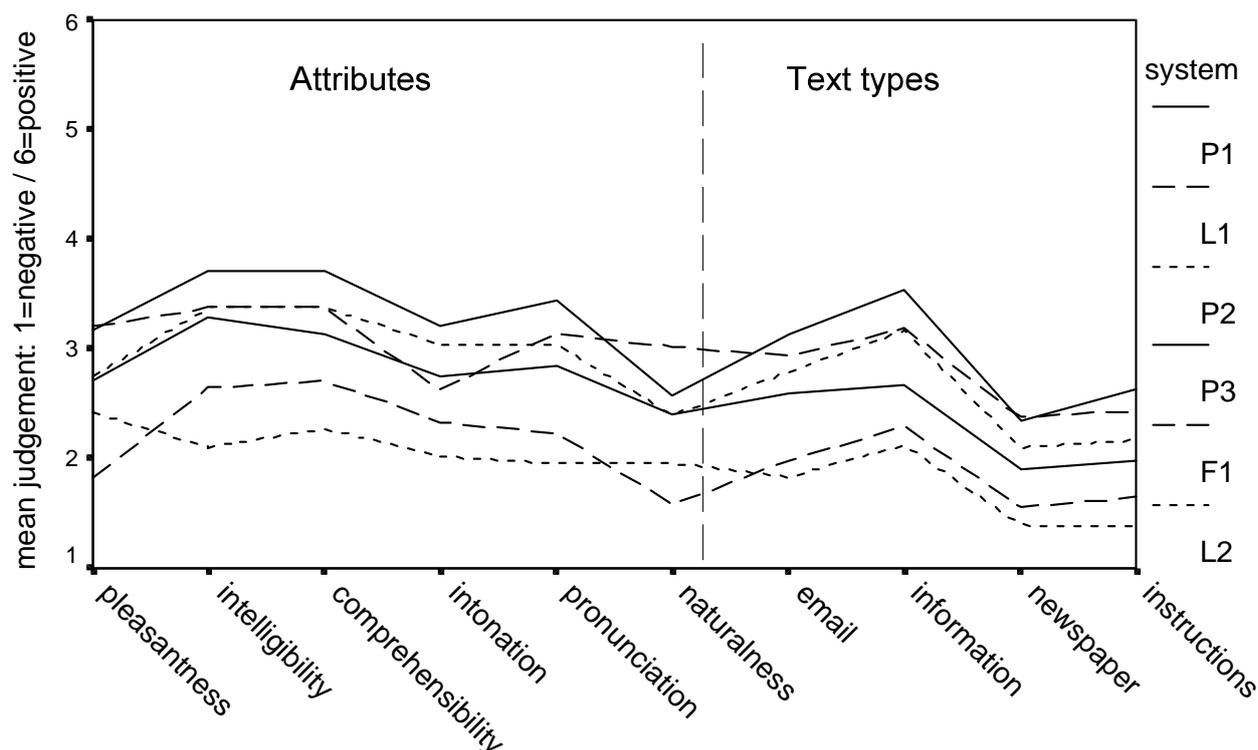
Figure 1: Mean rating for each question (see Table 2) across systems.

similar to the semantic differential paradigm; two extreme values (very good - very bad) served as anchors. It was assumed that by forcing the subjects to decide on one half of the scale maybe some

otherwise hidden biases could be revealed. 24 students (14 female, 10 male) gave 10 ratings on each text while listening to it. 6 ratings concerned general voice attributes and 4 were application specific questions (Table 2).

All the subjects were paid for their participation. The test sessions lasted around 45 minutes, took place in a quiet room, and was entirely computer guided. Presentation order was balanced across subjects. After rating one text read by each voice, the subjects were presented a single sentence from the text (one with no text conversion errors). According to this one sentence presented in all six voices, subjects were asked to assign a general ranking order. It was assumed that after hearing each voice reading the whole text, the subjects would recognize the voice and base their ranking order decision on the voices in general and not on the one sentence only. At the end of the test session each subject answered a multiple choice question asking for the most annoying thing about the synthetic voices in general: problems concerning a) intonation, b) voice quality or c) text conversion.

---

1. How pleasant is the voice?
2. How hard is it to understand single sounds/words?
3. How hard would it be to grasp the message if you listen to it for the first time?
4. How do you judge mistakes in accentuation?
5. How do you judge the pronunciation?
6. How natural do you judge the voice?

How would you like this voice to read to you on a daily basis
7. - your email or fax messages via mobile phone?
8. - a telephone information (e.g. timetables)?
9. - the news from a paper whilst you're driving your car?
10. - the instructions of a complicated telephone?

Table 2: Rated questions: 1-6 concerning voice quality, 7-10 concerning specific applications.

4. RESULTS

4.1 Rated attributes

Differences in system ratings as described in Figure 1 were significant (p<0.01) for each rated attribute. However, the differences cannot be directly related to one specific attribute. A Tamhane-Post-Hoc comparison showed the significant differences between the individual systems (table 3). The average value is 2.6, which is remarkably lower than the arithmetic mean of 3.5, thus indicating that the listeners were not satisfied with the quality of the TTS systems. Even the best system scored below 3.5 in nearly all categories.

System	P1	L1	P2	P3	F1	L2
P1		-	x	x	x	x
L1	-		-	x	x	x
P2	x	-		-	x	x
P3	x	x	-		x	x
F1	x	x	x	x		-
L2	x	x	x	x	-	

Table. 3: Significant differences between the individual systems with averaging over all categories (x = significant).

Although the results for the different attributes seem to be similar the pairwise correlation coefficients between them are only around 0.6 (0.4 < cc < 0.7). A factor analysis resulted in one main factor explaining 62% of the variance (in contrast to two factors described in [2], where a similär test was carried out and the factor analysis revealed two main factors that can be interpreted as intelligibility and naturalness).
It can be seen that all ratings are rather low, and that non-parametric systems deliver speech that is judged better than the speech produced by other systems (with L1 being one notable exception).

4.2 Ranking order

The results of the ranking order question are displayed in Figure 2. The median of the ranking order judgements was calculated. It seems that three "good" and three "bad" systems can be distinguished, and that ranking in these groups

differs due to personal preferences but also due to the preceding text.. However, the results of both questions are fairly similar to each other and to the results obtained with the category estimation questions.

4.3 General 'nuisance factor'

Prosody (43 %) and voice quality (40 %) were the factors that contributed most to the bad overall judgement of the synthetic speech quality, while text processing (17 %) was mentioned comparably seldom.

5. DISCUSSION

The most important result is that current TTS systems do not deliver the quality that customers demand. They sound unnatural and their speech is not pleasant to listen to. Neither system scored higher than 3.0 on any application (email reading, newspaper reading, technical manual) which is clearly on the lower half of the scale. Intelligibility was judged best which is necessary but obviously not sufficient for success. A recent evaluation under GSM conditions with similar voices [3] indicates that intelligibility is in fact quite high for most systems, which is reflected by the fact that, contrary to earlier investigations with less mature systems [2,6], a factor analysis did not yield two factors (one associated with naturalness, and the other with intelligibility).
But intelligibility is not everything. The quality of the prosodic realization and the overall voice quality seem to be the current challenges for speech synthesis developers while errors in text processing are less important (and quite infrequent anyway). However, the best system P1 scored last in [5] where due to a sophisticated delexicalization procedure only the prosody was evaluated. Pleasantness and acceptability are complex perceptual impressions that are influenced not only by the quality of the single components of a synthesis system like text processing, prosody control etc., but also by a well balanced interplay between them.
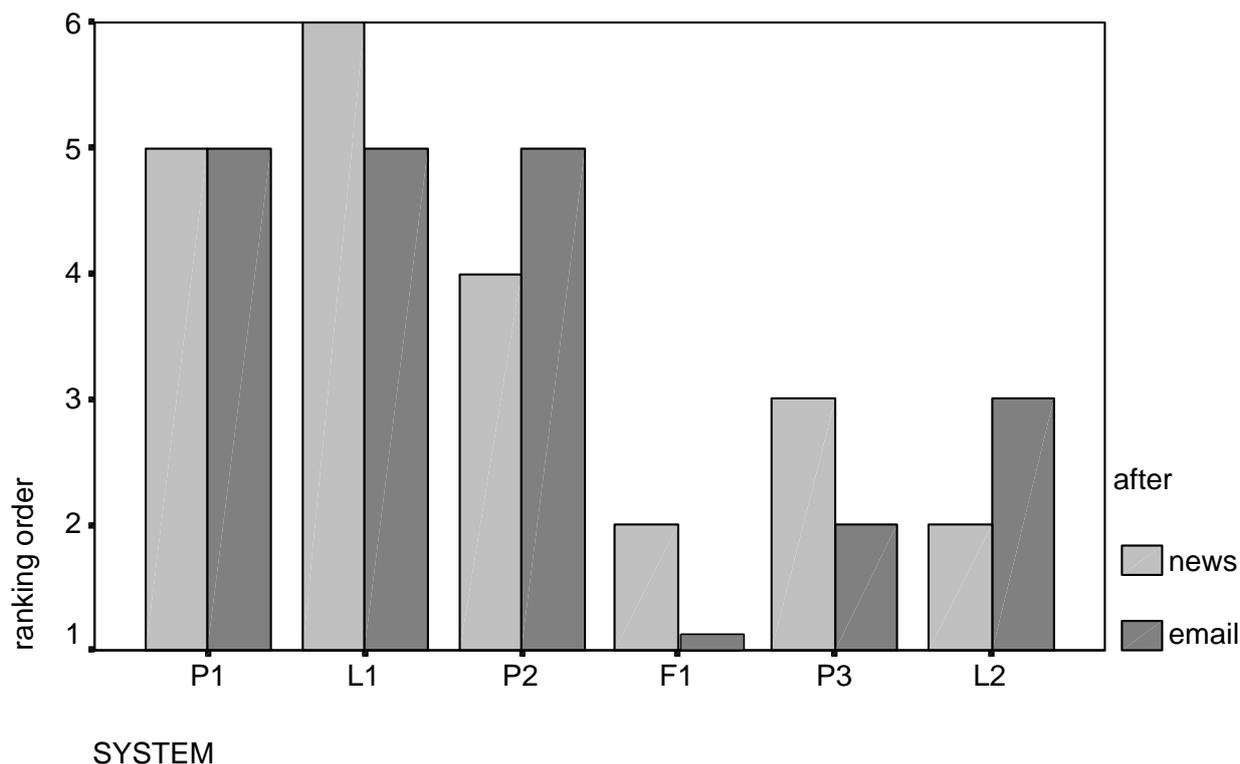
Figure 3: Median of ranking order position for each system, according to the time of judgement.

6. CONCLUSION

It seems that current speech synthesis methods have not reached a level where general customer satisfaction can be expected or even guaranteed. It remains to be seen whether new techniques like[7,8,9] will be able to "deliver on the long-standing promise of truly natural-sounding speech" [8].

7. ACKNOWLEDGEMENTS

REFERENCES

[1] Silverman,K.; Basson,S.; Levas,S. (1990) "Evaluating synthesiser performance: is segmental intelligibility enough?", in: Proceedings of the International Conference on Spoken Language Processing (ICSLP), 981-984

[2] Kraft,V.; Portele,T. (1995) "Quality Evaluation of Five German Speech Synthesis Systems", in: Acta Acustica, vol.3, 351-365

[3] Sonntag, G.P.; Portele, T.; Haas, F. (1998) "Comparing the comprehensibility of different synthetic voices in a dual task experiment", in: Proc. 3rd ESCA Workshop on Speech Synthesis, Jenolan Caves, 5-10

[4] ITU-T (1993) "A method for subjective performance assessment of the quality of speech voice output devices", Draft ITU-T Recommendation P.85, COM 12-R 6

[5] Sonntag,G.P.; Portele,T. (1998a) "Comparative evaluation of synthetic prosody with the PURR method", in: Proc. ICSLP, Sydney, Australia

[6] Klaus,H.; Fellbaum,K. (1997) "Auditive Bestimmung und Vergleich der Sprachqualität von Sprachsynthesesystemen", in: ACUSTICA/Acta Acustica, vol.83, 124-136

[7] Campbell, N (1999) "Data-driven speech synthesis", in: Proc. Forum Acusticum 1999, Berlin, S103

[8] Beutnagel, M.C.; Conkie, A.D.; Schroeter, J.; Stylianou, Y.; Syrdal, A.K. (1999): "The AT&T Next-Gen TTS System", in: Proc. Forum Acusticum 1999, Berlin, S104

[9] Stöber, K.; Portele, T.; Wagner, P.; Hess, W. (1999): "Synthesis by Word Concatenation", this volume