A User Study on the Influence of Mobile Device Class, Synthesis Method, Data Rate and Lexicon on Speech Synthesis Quality

Michael Pucher, Peter Fröhlich

ftw. Telecommunications Research Center Vienna, Austria
pucher@ftw.at, froehlich@ftw.at

Abstract

In this paper, we report on a comparative user study about the quality of mobile speech synthesis methods. We measured the impact of device class, data rate, synthesis method (diphone vs. non-uniform unit-selection) and lexicon usage on speech quality (word comprehension and several subjective satisfaction metrics). Seven practically relevant speech synthesis implementations and one natural voice were evaluated, applying the method recommended in ITU-T P.85, with additional pairwise comparisons.
As a general result, although the overall subjective ratings of the synthetic voices differed significantly, the word comprehension rates were quite similar. We found a significant impact of data rate and synthesis method on the mean subjective speech quality, but not on word comprehension. The use of a lexicon in embedded speech synthesis slightly improved the perceived pronunciation quality.

1. Introduction

1.1. Background

Speech interaction offers much potential in many mobile usage situations serving as a useful compensation for tedious gestural input and strenuous visual display facilities of mobile devices. Based on the strong need to find a coherent framework for the integration of speech, visual display and key and pen gesture control into a multitude of mobile target devices, our project MONA [1] developed a multimodal presentation server and two multimodal mobile demonstrators.
When integrating speech in mobile applications, several constraints have to be considered. Since for embedded speech synthesis today's mobile devices set a limit on the database size, a diphone-based synthesis method has to be applied. Non-uniform unit-selection synthesis methods, using large databases are only available with a server-based approach. When dealing with server-based synthesis it is essential which data-rate/codec is used. We were interested in finding more empirical evidence about the effects of these implementation issues on user perception.
The major evaluation criteria of text-to-speech (TTS) quality are intelligibility (or comprehension) and naturalness of speech [2]. Speech synthesis products for personal computers have already achieved a high intelligibility since several years, with, objective word intelligibility scores for the best TTS systems of up to 97% [2] [3]. TTS evaluation studies concerning naturalness and other subjective quality criteria like "overall voice quality" show that speech synthesis systems are improving to a satisfactory degree, but still are not comparable to the natural human voice [4].

The ITU-T P.85 standard [5] proposes a methodological framework for a combined evaluation of speech intelligibility and naturalness, including a concrete procedure and rating scales. A reliability evaluation of this standard showed that it gives identical rankings but a higher variation than pairwise comparison alone [4]. In order to make our study replicable and its results comparable, we applied this methodological framework to our study.

1.2. Research questions

Relevant research questions resulting from these considerations are:

1. How do the investigated synthetic sources perform in general regarding comprehension and speech quality, compared to natural human speech?
2. What is the influence of the device class on the speech synthesis quality?
3. Do the more advanced synthesis capabilities of the server-based approach lead to better speech quality evaluation results?
4. Can speech synthesis quality be increased when using a user-defined lexicon?
5. In the case of server-based speech synthesis, in which way do different data rates/codecs influence speech synthesis quality?

2. Method

Keeping very close to the recommendations given by the ITU-T P.85 [5] standard, the experiment was a balanced, within-subjects test-design. The independent variables were the different sources (see section 2.2). The speech material was selected from the two mobile demonstrators within our project [1], see section 2.3.
Dependent variables were word comprehension (see section 2.4.1.), and the P.85 naturalness rating scales (see section 2.4.2.). As a further dependent variable, which is not yet included in the standard, but which has proven to be particularly reliable [4], we also included pairwise comparisons (see section 2.4.3.). The TTS technology used in the evaluation was provided by SVOX [6].

2.1. Participants

15 paid users recruited by web- and radio-based announcements took part in the study. Regarding the broad target population of mobile speech-enabled applications, major demographic variables, such as gender, age (below and over 30 years) and professional status were balanced. The majority of test users (12) were not familiar with synthesized speech.

2.2. Sources

Our variables of interest - device class, synthesis method, data rate and lexicon – were operationalized as follows:
As *device classes* we chose a typical SmartPhone (device class 2 according to [7]; Nokia 7650/Symbian) and a PocketPC (device class 3 according to [7]; PocketPC 2003/WinCE). Concerning *synthesis method*, both the SmartPhone and Pocket PC were equipped with an embedded TTS engine. The server-based TTS was only used on the Pocket PC. The *data-rates/codecs* of interest for our study were a 128 kbps PCM , a 13 kbps GSM, and a 4,75 kbps AMR codec.
Embedded speech synthesis versions with a *lexicon*, containing otherwise wrongly pronounced words, such as proper names and foreign words, were compared to a non-lexicon version.
For the experiment, we chose 7 sources, i.e. combinations of the above-described constituents of mobile TTS processing. In addition, as a control condition, we included a natural voice of a female speaker played back over the Pocket PC (see Table 1 below).

Nr.	Source-Name	Device class	Synth. method	Data-rate/ Codec	Lex
1	Smart_NoLex	Smartphone	Emb.	16kHz	No
2	Smart_Lex	Smartphone	Emb.	16kHz	Yes
3	Pocket_Emb_NoLex	Pocket PC	Emb.	8kHz	No
4	Pocket_Emb_Lex	Pocket PC	Emb.	8kHz	Yes
5	Pocket_Serv_4.75	Pocket PC	Server	8kHz 4.75 kbps AMR	Yes
6	Pocket_Serv_13	Pocket PC	Server	8kHz 13 kbps GSM	Yes
7	Pocket_Serv_128	Pocket PC	Server	8kHz 128 kbps PCM	Yes
8	Natural	Pocket PC	%	16kHz	%

Table 1: Used Sources

In order to reduce the number of sources included in the experimental design, we did not combine each of the four constituents with each other. Choosing a pragmatic approach, we only included those combinations that would actually be established in practice and would therefore enable us to achieve ecologically valid results. For example, a full combinatorial experimental design would, in theory, require us to evaluate the combination of the server-based voices with the non-lexicon condition. However, since in the server-based approach there are no storage constraints, this solution would never be implemented in practice.

2.3. Speech material

As recommended by the P.85 standard, we used text messages from two different domains (4 each, thus 8 in total). These were the two applications developed in our project [1]: an Email client (MONA@work) and a multiparty chat game (MONA@play). The speech material for the MONA@play application were questions and their corresponding answers ("Who won the soccer world cup in 1978? Mexico, Germany").
For MONA@work it was an Email header consisting of Date, Sender and Subject ("10. June 2003 from Bogusch Claudia. Date for the barbecue.").
These tasks contained a variable and a fixed part according to the P.85 recommendation. However, in case of the questions, the fixed part is only the question/answer structure. Additionally we had two training tasks, which were not included in the analysis.

2.4. Procedure

The experimental study was conducted in individual test sessions, taking about 80 minutes.
In an introductory part, subjects were interviewed to gather basic demographic data and previous knowledge about the speech synthesis domain. They were introduced into the test contents and procedure.
During the test, each of the 8 spoken messages was presented to the user by one of the 8 speech sources described above. Speech messages and sources were combined in a way that each message-source combination was evaluated throughout the study. Each subject was exposed to each message and each source once. The subjects entered their responses into a set of web forms we designed containing the translated evaluation forms provided in [5].

2.4.1. Comprehension Test

Each speech message was presented twice. After the first presentation, the subjects were asked to enter the content of the spoken message into a text field provided on the web form. For the MONA@play messages, both the reproduction of the question and the answer was required. For the MONA@work messages, we asked the subjects to enter the date, time, sender, and the subject.

2.4.2. Subjective Speech Quality Ratings

After the second presentation of each message, the subjects judged the subjective speech quality by expressing their opinion on the rating scales provided in [5]. There were several 5-point rating scales dealing with sub-aspects of voice quality judgment:

1. Overall Impression: "How do you rate the quality of the sound of what you have just heard?"
2. Listening Effort: "How would you describe the effort you were required to make in order to understand the message?"
3. Comprehension Problems: "Did you find certain words hard to understand?"
4. Articulation: "Were the sounds distinguishable?"
5. Pronunciation: "Did you notice any anomalies in pronunciation?"
6. Voice pleasantness: "How would you describe the voice?"

Another quality judgment scale was about acceptance, only offering "yes-no" answering options: "Do you think that this voice could be used for such an information service by telephone?" A further scale aimed at assessing the speaking rate: "The average speed of delivery was …? ".

2.4.3. Pairwise comparisons

Finally, after all sources had gone through comparison testing and subjective voice quality ratings as described above, we also asked the subjects to make pairwise comparisons of the sources using a single message. We asked the users to decide which voice they liked more, concerning overall quality. To keep the number of pairs low, we excluded 3 sources *Pocket_Serv_13*, *Pocket_Emb_NoLex* and *Smart_NoLex*,. With the remaining 5 pairs, a comparison of high and low data rate source and all sources with lexicon was possible.

3. Results

We will first give an overall impression of the study results (answering research question 1), then we will specifically address the influence of the defined independent variables on word comprehension and subjective quality rating (research questions 2-5).

3.1. Overall Comparison

In order to analyze the *word comprehension* results, we defined a word error rate measure (P.85 does not provide an explicit recommendation in this regard).
Figure 1 shows the mean word comprehension rates for the 8 different sources, including 95% confidence intervals.

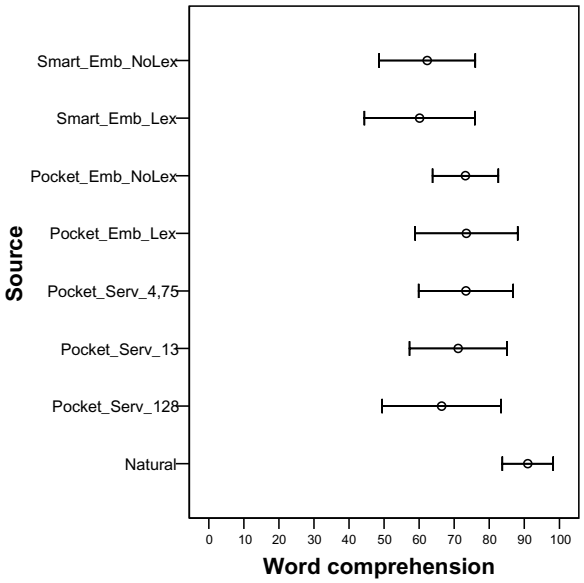

Figure 1: Comprehension of Words

The mean comprehension rates of the synthesis methods ranged between 60% (for the SmartPhone solutions) and 72% (for the embedded Pocket PC solutions and for the server-based 4.75 data rate solution). The sources did not significantly differ from each other (Wilcoxon-tests for paired samples).
However, the word comprehension rate of the natural voice was significantly higher than all synthetic voices (mean word comprehension rate: 95%, Wilcoxon-test, p<.05).
All mentioned rating scales dealing with sub aspects of *subjective voice quality* correlated highly significantly (p<.01), between r=.40 for listening effort and voice pleasantness and r=.80 for listening effort and comprehension problems. In order to get an overall picture of the subjective quality ratings, we computed a "Mean satisfaction" scale out of scales 1-6 (see Figure 2).

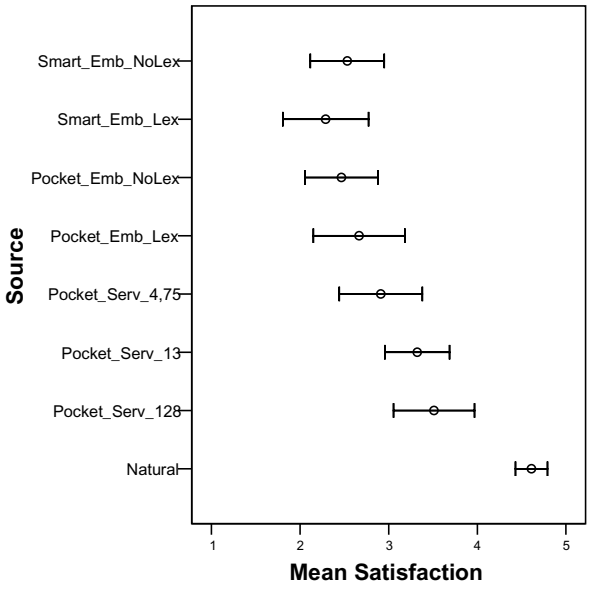

Figure 2: Mean Satisfaction

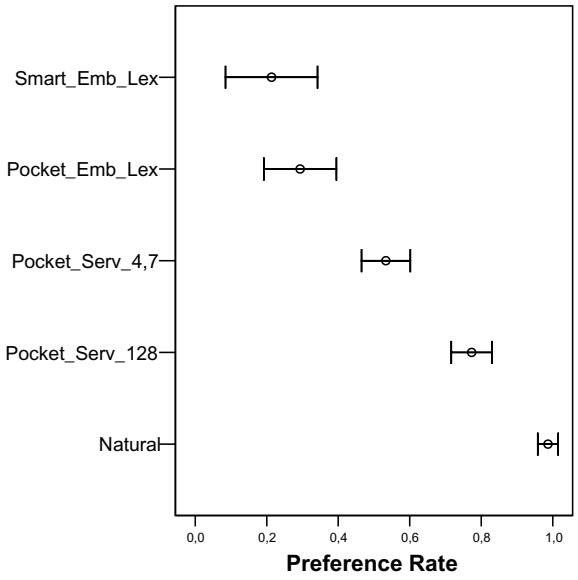

Figure 3: Pairwise Comparison

Similarly to the comprehension rates, the mean satisfaction results of the natural voice were significantly better than all synthetic voices (Wilcoxon-tests, p<.001). However, it is obvious that there are now stronger differences between the synthetic sources.
During the pairwise comparison session (see Figure 3), the listeners preferred the server-based sources significantly more often than the embedded versions (Wilcoxon-tests, p<.01).

The speaking rate was generally found to be nearly optimal, in some cases slightly too slow. However, individual users were quite consistent in their speaking rate judgments across voices. [8] argues with confirmatory factor analysis, that the speaking rate belongs to the intelligibility factor.

3.2. Device class (for embedded synthesis)

No significant differences between word comprehension or subjective quality ratings between the embedded SmartPhone sources (No 1 and 2) and Pocket PC sources (No 3 and 4) could be found.

3.3. Synthesis Method

The server-based sources (5–7) were rated significantly better than the embedded sources on the PocketPC (3,4). This concerned all subjective quality ratings (Wilcoxon-tests, $p<.05$, mean differences between 0,6 and 1 rating points), except "Voice pleasantness". However, no significant differences in word comprehension were obtained.

3.4. Data rate/Codec

A general trend was that the higher data rate of the 128 kbps PCM coded server-based voice was rated best, followed by the lower data rate codecs. The Wilcoxon test results indicated that the data rate particularly influenced the perceived pleasantness of the voice, and the acceptance rate. There were also differences concerning *subjective* comprehension problems, although there no significant differences were found in the *actual* word comprehension. In the relation between *Pocket_Serv_4.75 (AMR)* and *Pocket_Serv_13 (GSM)* there is also a highly significant difference concerning the overall impression. This is interesting, since one would suspect that the difference between *Pocket_Serv_4.75 (AMR)* and *Pocket_Serv_128 (PCM)* is the most significant.

3.5. Lexicon

For the pronunciation rating scale, we found a significantly better rating during the lexicon-condition compared to the non-lexicon condition (Wilcoxon-test, $p<.05$). However, this effect was small (only 0.5 scale points) and it only applied for the Pocket PC (i.e. sources 1 and 2), not for the SmartPhone (sources 3 and 4). The question for pronunciation was: "Did you notice any anomalies in pronunciation". We can conclude that it is possible to correct pronunciation anomalies by using a lexicon, although this does not increase the voice quality significantly on other scales.

In a short informal comparison of device loudspeakers with headphones during the test, no systematic differences were found.

4. Conclusions

Our results show that although the subjective voice quality ratings differed significantly, the comprehension of words is very similar. Although similar results were already reported in [3], this is still somewhat surprising, since our systems were varied in many respects. Concerning the use of synthetic speech in mobile applications one should consider this difference and use any system when mostly the comprehension is important and use a high quality system when the subjective voice quality ratings are important. The latter is especially relevant in the domain of entertainment applications, like in MONA@play.

Considering the overall subjective quality ratings one has to take into account, that most users were not familiar with speech synthesis, and that all speech messages included proper names, which are especially difficult for speech synthesis.

We also saw that for certain embedded voices pronunciation anomalies can be corrected using a user defined lexicon. The data rate for the server based voices was also significant, and especially here we saw a division between the objective comprehension, measured by word error rate, and the subjective voice quality rating concerning comprehension problems. Users reported comprehension problems, although they did not significantly influence the objective comprehension of words.

Concerning the ITU-T P.85 standard it could be considered to extend it with guidelines for defining sources, source-task combinations, and to develop word error rate metrics, which are not yet provided by the standard.

5. Acknowledgements

We wish to thank SVOX for contributing their TTS technology that made this study possible. The MONA project was funded by Kapsch CarrierCom AG, Mobilkom Austria AG and Siemens Österreich AG together with the Austrian competence centre programme Kplus. Many thanks to Elisabeth Muss for preparing the data and defining and calculating the word error rate for word comprehension.

6. References

[1] MONA – Mobile Multimodal Next Generation Applications, http://mona.ftw.at/

[2] Kamm, C., Walker, M. and Rabiner, L. "The Role of Speech Processing in Human-Computer Intelligent Communication", *Speech Communication, Volume 23, Issue 4, p263-278, 1997.*

[3] Beutnagel, M., Conkie, A., Schroeter, J., Stylianou Y. and Syrdal, A. "The AT&T Next-Gen TTS System". *Joint Meeting of ASA, EAA, and DAGA, Berlin, Germany, March, 1999.*

[4] Alvarez Y. V. and Huckvale M. "The Reliability of the P.85 Standard for the Evaluation of Text-to-Speech Systems", *ICSLP 2002, Denver, Colorado.*

[5] ITU-T P.85 "A method for subjective performance assessment of the quality of speech voice output devices", *06/1994.*

[6] SVOX Ltd., *http://www.svox.com*

[7] Giller V., Melcher R., Schrammel J., Sefelin R., and Tscheligi M. "Usability Evaluations for Multi-device Application Development Three Example Studies", *Human-Computer Interaction with Mobile Devices and Services, LNCS 2795, p302-316, 2003*

[8] Viswanathan M., Viswanathan M. "Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale", *Computer, Speech and Language 19 (2005), 55-83*