On the Use of Automatic Speech Recognizers for the Quality and Intelligibility Prediction of Synthetic Speech

Florian Hinterleitner, Steffen Zander, Klaus-Peter Engelbrecht, Sebastian Möller

Quality and Usability Lab, TU Berlin, Germany
florian.hinterleitner@tu-berlin.de

Abstract: In this paper we investigate the use of an automatic speech recognizer (Google Speech API) for the prediction of quality and intelligibility of synthetic speech. For two databases of rated synthetic speech samples, we analyze the correlation of the word error rates (WER) obtained from the recognizer for each sample with ratings on 16 different attribute scales. Moderate correlations are observed for various quality aspects including *overall impression*, *naturalnesss*, and *intelligibililty*. Moreover, we analyze in a third database the correlation between intelligibility by a human, as determined in a test with semantically unpredictable sentences, and the WER of the recognizer. The correlation between the humans' and the recognizer's WER over all samples is .40, and .94 if averaged by TTS system.

1 Introduction

The quality of synthetic speech has increased immensely over the past years. Nowadays, Text-to-Speech (TTS) systems have attained a level of quality that makes it feasible to integrate them into everyday services like email readers, information systems, and smart home assistants. An increasing number of people get in contact with synthetic speech every day due to, e.g., Apple's Siri or the popularity of e-books and the implied possibility to synthesize a book's content. With further applications emerging on the market high quality TTS systems become even more important. During the development such systems are usually assessed through listening tests to evaluate the quality perceived by the user. The most common method comes in the form of a semantic differential where test participants rate stimuli on attribute scales, e.g., natural vs. unnatural [1]. Such a test protocol is standardized in the ITU-T Rec. P.85 [2]. Apart from a semantic differential TTS signals can also be assessed via a paired comparison test with subsequent multidimensional scaling [3]. In these tests participants are asked to rate the similarity within pairs of TTS stimuli. Thus, the quality rating of a participant solely depends on their perceptual quality impression and is not biased by any given ratings.
However, the both approaches for quality assessment have in common that they are usually a cost-intensive and time consuming process. To reduce the amount of such tests an objective measure that identifies weaknesses of a system in a multidimensional quality space would be of great benefit for developers. Numerous approaches have tackled this problem and achieved correlations between objective scores (i.e. the predicted scores) and subjective ratings of over .80 on multiple perceptual dimensions [4, 5].
In a previous study the usefulness of HMMs for the quality prediction of synthetic speech [6] was examined. The authors trained a HMM on natural speech and used it to measure the similarity of synthetic speech with the natural reference speech data that was used for the training. Even though the study showed the potential of HMMs for such a task the approach was limited due to the small natural speech training data (ca. 1 hour). Ideally such an acoustic model should be trained on a large-scale dataset. Since representative data of this dimension is hard to obtain,

especially for the German language, we examine the use of acoustic models of automatic speech recognizers (ASR) like the Google Speech API[1].
In this paper we investigate the benefit of ASRs for the quality and intelligibility prediction of synthetic speech. To do so, we first use the Google Speech API to recognize the content of speech files generated by different German and English TTS systems. Then, the resulting word error rate (WER) is used as a predictor for overall quality, intelligibility and various other quality indicators that were assessed in listening tests.
In Section 2 we describe the TTS databases that were used in this study. Followed by Section 3 in which we present and discuss the results. Finally, Section 4 concludes this study and gives an outlook to future work.

2 Databases

This section gives an overview of the 3 subjectively evaluated TTS databases (DB) that were used in this study.

2.1 Database 1

DB 1 was part of an extensive study by Hinterleitner et al. [1] in which perceptual quality dimensions of state-of-the-art TTS systems were investigated. 16 German-speaking synthesizers (formant synthesizers, PSOLA-based diphone synthesizers, unit-selection systems, and HMM-synthesizers) were used to generate 2 samples for each of the 30 different configurations of synthesizer and voice. The average duration was 10s. All stimuli were rated by 30 participants on 16 continuous attribute scales that were developed during two extensive pretests. These items can be linked to perceptual qualtiy dimensions such as naturalness of voice, prosodic quality, fluency & intelligibility, absence of disturbances, and calmness.

2.2 Database 2

This database was gathered during a subsequent study by Hinterleitner et al. [3] which aimed to complement and to expand the results from the study mentioned in the previous paragraph (Section 2.1). Therefore, 30 female and 27 male stimuli with an average duration of 5s were generated from one utterance by different configurations of German-speaking TTS systems (formant synthesizers, PSOLA-based diphone synthesizers, unit-selection systems, and HMM-synthesizers). The stimuli were evaluated by 40 naïve test participants in a sorting task with subsequent multidimensional scaling. In a post-test, all stimuli were also rated on the same 16 continuous scales as described in DB 1. 12 test participants (5 expert listeners from the Quality and Usability Lab of the TU Berlin and 7 naïve subjects) took part in the test. For this study we focused on the subjective ratings on the 16 attribute scales.

2.3 Database 3

DB 3 database was conducted during the Blizzard Challenge 2011 [7] which is an annual competition of developers of TTS systems. Since 2005 the Blizzard Challenge gathers developers of TTS systems to compare techniques in building corpus-based speech synthesizers. The fact that all participants get the same speech corpus to build their systems on assures a comparability between all synthesizers. In an extensive online evaluation different quality aspects like naturalness and similarity to the original voice are evaluated. These tests also feature semantically

[1]Google Speech API: http://www.google.com/intl/en/chrome/demos/speech.html

unpredictable sentences (SUS) to assess the intelligibility of each voice. For our research we are using the results of the SUS evaluation in the EH1 task in the Blizzard Challenge 2011. This database contains 26 stimuli for each of the 12 different TTS systems plus 26 stimuli of the original natural speaker (338 stimuli with an average duration of 2s.

3 Results and discussion

We used the Google Speech API to compute WERs for all stimuli in the databases DB 1-3. Google Speech API allows to upload a wav-file containing speech and returns a recognizer hypothesis (and a confidence value). From the hypothesis and the reference transcription of the sentence, we can compute the WER. To do so, we first normalize the strings. Then, we calculate the Levenshtein distance to obtain the minimal number of edits to transform one sequence of words into the other. The number of edits is divided by the total number of words in the reference transcription.
For DB 1, Google Speech API returned an error of unknown type for 20 of 60 samples. Thus, these stimuli had to be omitted from the analysis. Of the other databases, all samples could be processed.
The WER obtained from Google Speech API can be used to predict measures of the TTS quality and performance. We use the data from DB 1 and 2 to analyze correlations of the WER with the subjective quality ratings in several dimensions. Subsequently, we use data from DB 3 to compare the objective WER obtained from the recognizer with the subjective WER computed from the subjects' recognition performance. The following subsections show the results for all 3 databases.

3.1 Predicting TTS quality

Since the WERs were not normally distributed and included an outlier (in DB 1), we computed the Spearman rank-order correlation ρ_s between the WER and the attribute scales of DB 1 and 2. The highest correlations can be seen in Table 1.

Table 1 - Absolute correlation values between the attribute scales of DB 1 and 2 and the objective WER.

	DB 1 ($N = 40$)		DB 2 ($N = 57$)	
scales	$\lvert\rho_s\rvert$	p	$\lvert\rho_s\rvert$	p
OVERALL IMPRESSION (OI)	.35	$p < .05$	.63	$p < .001$
VOICE PLEASANTNESS (VP)	.30		.67	$p < .001$
DISTORTIONS (DI)	.31		.64	$p < .001$
CLINK (CL)	.31		.61	$p < .001$
NATURALNESS (NAT)	.26		.65	$p < .001$
INTELLIGIBILITY (INT)	.27		.61	$p < .001$
STRESS (STR)	.28		.55	$p < .001$
NATURAL RHYTHM (NATR)	.22		.56	$p < .001$
TENSION (TEN)	.22		.56	$p < .001$

Note: The table only features scales with an average $\lvert\rho_s\rvert$ between both databases of over .50.

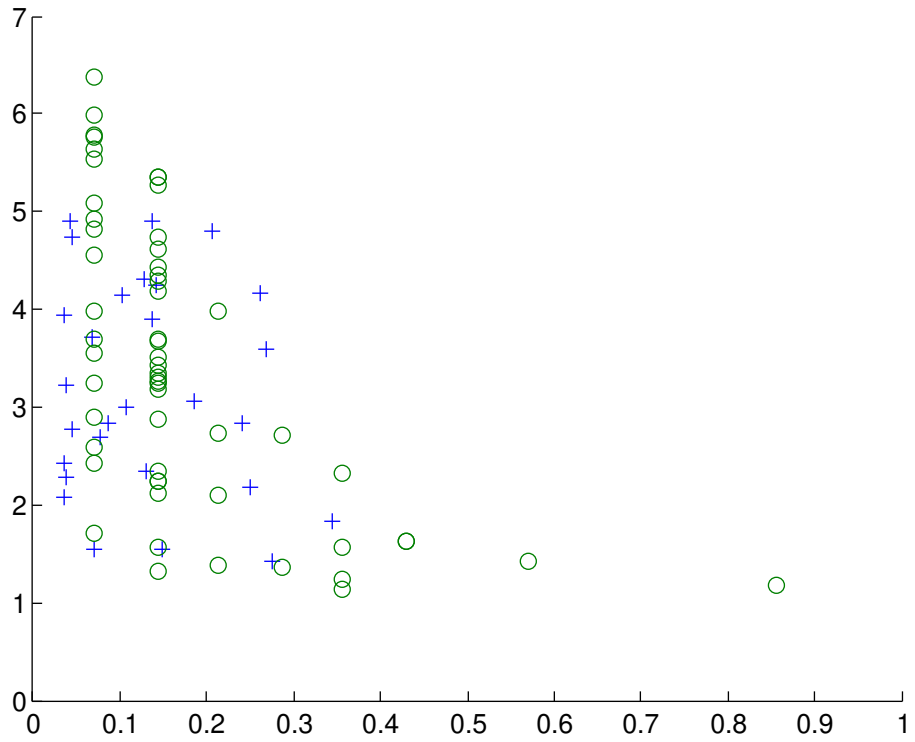

Figure 1 - Scatter plot of WER and overall impression ratings for DB 1 and DB 2. (Note: blue crosses mark the data of DB 1, and green circles mark data of DB 2.

The table clearly shows that the correlations in DB 2 are far higher than in DB 1 for all scales. Also, while all correlations in DB 2 in this table are highly significant ($p < .001$) the only scale of DB 1 that features a significant correlation is *overall impression* ($p < .05$). The scatter plot (Figure 1) shows the distribution of the data points for the overall impression rating in both databases. One reason for the higher correlation in DB 2 may be the larger spread in the WERs as well as the ratings.

Taking a look at the results for DB 2 shows correlations with 6 scales of over .60. These scales can be linked to perceptual constructs [8] as *naturalness of voice* (OI, VP, NAT), *disturbances* (DI, CL), and *fluency & intelligibility* (INT). The items with the lowest correlations for DB 2 can be linked to perceptual dimensions like *calmness* (item: speed), *disturbances* (items: noise, hissing), and *fluency & intelligibility* (items: fluency, bumpiness). Interestingly, in the dimensions *disturbances* and *fluency & intelligibility* we found very high and very low correlations. Thus, even items linked to the same perceptual quality dimension can feature different aspects of this dimension and therefor they can correlate different than other items linked to the same perceptual dimension.

3.2 Predicting TTS intelligibility

Considering the correlation of the WER with the intelligibility in Table 1 and that ASRs are used for the intelligibility evaluation of speech we conducted further research in this direction. Therefor, we chose the data of the SUS test from the Blizzard Challenge 2011 (DB 3) and computed the Spearman rank-order correlation r_s between the WER by the recognizer (objective WER) and the WER computed from the subject's transcriptions (subjective WER). We also computed the correlation between the mean WER for each synthesizer of the challenge and their corresponding mean objective WERs. The results are shown in Table 2.

Table 2 - Correlations between subjective and objective WER for DB 3

	r_s	rmse
INTELLIGIBILITY PER FILE	.40	0.24
INTELLIGIBILITY PER SYSTEM	.94	0.15

Note: p-values for all correlations $< .001$.

While the correlation per file only reaches a moderate level, the correlation of the WER scores averaged by system comes close to 1. The scatter plot in Figure 2 further illustrates this relationship. Even though the correlation is extremely high you can also see that the WER detected by the Google Speech API is on average higher than the subjective WER. Thus, even though an intelligibility ranking of systems via a speech recognizer seems feasible the WER scores are generally overestimated.
Considering the very small range of subjective WER (0.12 to 0.21) when compared on a per system basis, the very high correlation is even more outstanding since many systems on a similar intelligibility level result in a greater challenge for a predictor.

4 Conclusions and future work

In this study the Google Speech API was used to compute WERs for TTS samples in 3 different TTS databases. The stimuli of the first two databases were evaluated on 16 attribute scales in a listening test. In the third database, intelligibility by humans was evaluated in a test with semantically unpredictable sentences (SUS)
First, we computed the Spearman rank-order correlation between the WER obtained from the Google Speech API and the scores on the attribute scales of DB 1 and 2. The correlations for DB 1 only reached fairly low values. However, for the files of DB 2 we obtained correlations above .60 for 6 scales. This indicates that the WER produced by a speech recognizer holds valuable information concerning different perceptual quality dimensions of TTS systems like the *naturalness of voice* and *fluency & intelligibility*, as well as concerning the users' overall impression of the system.
Given the high correlations with the intelligibility scale, we used the TTS files of DB 3 which were evaluated in a SUS test concerning their intelligibility. The correlations between the subjective and the objective WER for all files was .40. The correlation between the WER scores averaged by TTS systems however reached .94. Even though the rmse of 0.15 and the generally somewhat overestimated WER suggests that the speech recognizer can not accurately predict the intelligibility score of a system, an intelligibility ranking of different systems seems feasible.
Further research will concentrate on reproducing this result on other TTS databases that were evaluated in a SUS test but feature a wider intelligibility range between the TTS systems.

5 Acknowledgements

The present study was carried out at Quality and Usability Lab, TU Berlin. It was supported by the Deutsche Forschungsgemeinschaft (DFG) and the Bundesministerium fï¿½r Bildung und

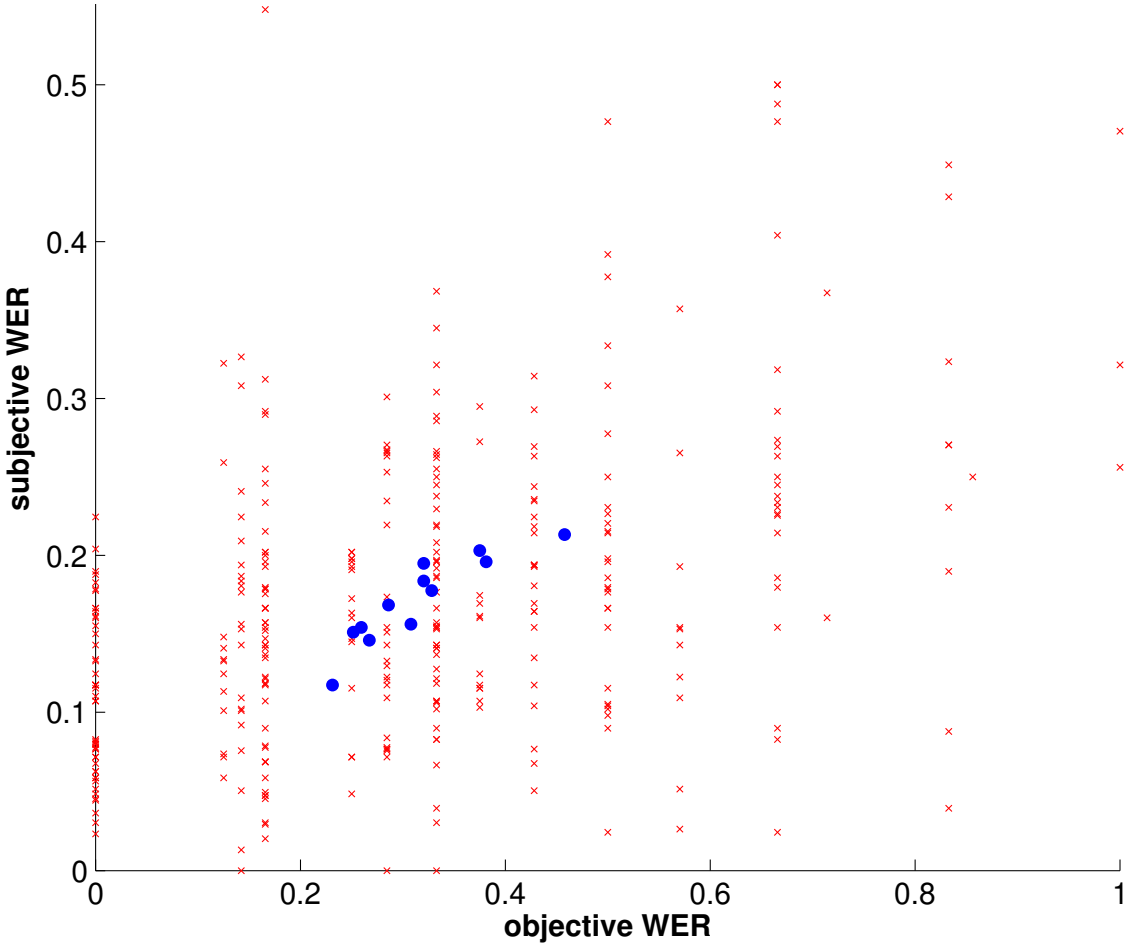

Figure 2 - Scatter plot of subjective and objective WER for DB 3. (Note: red crosses mark the values for each stimulus whereas blue dots denote the average values per system.

Forschung (BMBF), grants MO 1038/11-1, MO 1038/11-2, HE 4465/4-1, HE 4465/4-2, and 01IS12056.

References

[1] F. Hinterleitner, S. Möller, C. Norrenbrock, and U. Heute. Perceptual Quality Dimensions of Text-to-Speech Systems. *Proc. of the 12th Annual Conference of the International Speech Communication Association (Interspeech 2011)*, pages 2177–2180, 2011.

[2] ITU-T Rec. P.85. *A Method for Subjective Performance Assessment of the Quality of Speech Voice Output Devices*. International Telecommunication Union, Geneva, 1994.

[3] F. Hinterleitner, C. Norrenbrock, S. Möller, and U. Heute. What Makes this Voice Sound so Bad? A Multidimensional Analysis of State-of-the-Art Text-to-Speech Systems. *Proc. of the 2012 IEEE Workshop on Spoken Language Technology (SLT 2012)*, pages 240–245, 2012.

[4] F. Hinterleitner, C. Norrenbrock, S. Möller, and U. Heute. Predicting the Quality of Text-To-Speech Systems from a Large-Scale Feature Set. *Proc. of the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013)*, pages 383–387, 2013.

[5] C. Norrenbrock. *Instrumental Quality Estimation for Synthesized Speeh Signals*. PhD thesis, Christian-Albrechts-Universität zu Kiel, 2014.

[6] T. H. Falk and S. Möller. Towards Signal-Based Instrumental Quality Diagnosis for Text-to-Speech Systems. *IEEE Signal Processing Letters*, 15:781–784, 2008.

[7] S. King and V. Karaiskos. The Blizzard Challenge 2011. *Proc. of the Blizzard Challenge Workshop. International Speech Communication Association (ISCA)*, 2011.

[8] F. Hinterleitner, C. Norrenbrock, and S. Möller. Is Intelligibility Still the Main Problem? A Review of Perceptual Quality Dimensions of Synthetic Speech. *Proc. of the 8th ISCA Speech Synthesis Workshop (SSW 2013)*, pages 167–171, 2013.