Comparison of Approaches for Instrumentally Predicting the Quality of Text-To-Speech Systems

Sebastian Möller[1], *Florian Hinterleitner*[1], *Tiago H. Falk*[2], *Tim Polzehl*[1]

[1] Quality and Usability Lab, Deutsche Telekom Laboratories, TU Berlin, Germany
[2] Bloorview Research Institute, Toronto, Canada

sebastian.moeller@telekom.de, florian.hinterleitner@googlemail.com,
tiago.falk@ieee.org, tim.polzehl@telekom.de

Abstract

In this paper, we compare and combine different approaches for instrumentally predicting the perceived quality of Text-to-Speech systems. First, a log-likelihood is determined by comparing features extracted from the synthesized speech signal with features trained on natural speech. Second, parameters are extracted which capture quality-relevant degradations of the synthesized speech signal. Both approaches are combined and evaluated on three auditory test databases. The results show that auditory quality judgments can in many cases be predicted with a sufficiently high accuracy and reliability, but that there are considerable differences, mainly between male and female speech samples.

Index Terms: speech synthesis, quality prediction, Quality of Experience (QoE)

1. Introduction

Text-To-Speech (TTS) systems have reached a level of maturity which allows them to be used in every-day spoken dialogue applications where flexibility and unlimited vocabulary are necessary, such as email and SMS readers, traffic information systems, or smart-home assistants. Providers of such systems have to select one of a number of available TTS systems for inclusion in their service, or to justify whether an update of their voice has really led to improvements. In parallel, developers of TTS systems frequently have to test their system during the development cycle, in order to optimize algorithms and corpora. Thus, the evaluation of synthesized speech is still a frequent and important task.

So far, evaluators of TTS systems rely on auditory tests with human participants. Depending on which aspect of the system is under consideration, different types of test are recommended: articulation and intelligibility tests investigate whether the synthesized speech signal is able to carry information on a segmental or supra-segmental level [1]; comprehension tests investigate whether the content provided via a synthesized speech signal can be discerned [2]; and overall quality tests investigate global aspects of the synthesized speech signal in an application scenario, such as naturalness, pronunciation, intonation, speech rate, voice pleasantness, etc. [3]. Although doubts have been casted on the test protocol [4][5], the latter method is by far the most frequently applied one when it comes to judging the overall quality of synthesized speech to be integrated in spoken dialogue services; as a consequence, it is recommended by International Telecommunication Union for evaluating telephone-based services [3]. Common to all these methods is that they rely on perceiving and judging listeners, which makes them time-consuming and expensive.

In order to increase efficiency, several proposals have been made in the last two decades to estimate the perceived quality of synthesized speech signals in an instrumental (sometimes called "objective") way. For concatenative speech synthesis, one idea is to collect natural speech samples from the same speaker which is used for the synthesis inventory; a perceptually weighted distance between the synthesized and the naturally-produced samples of this speaker can then be used as an index of the quality degradation [6]. Whereas this approach is similar to legacy pattern-comparison approaches used for predicting the quality of transmitted speech [7][8], it is only rarely applicable, as the inventory speaker is usually not available to the evaluator.

Mariniak [9] proposed to extract perception-based features from the synthesized speech material and to compare them to features extracted from (other) natural speakers; the distance between both could be an indication of the speech quality. To our knowledge, this approach was never implemented by Mariniak, but it has recently been taken up in [10], using Mel-Frequency Cepstral Coefficients (MFCCs) as features and a Hidden Markov Model (HMM) with Gaussian Mixture densities for a temporal-spectral comparison of features. It led to very promising results on an initial test database, with correlations between 0.54 and 0.81 for different quality dimensions collected in the auditory test.

Another approach is to extract parameters from the speech signal which are related to degradations typically expected for TTS. Also this approach is motivated by quality prediction-models for transmitted natural speech, namely the single-ended model given in ITU-T Rec. P.563 [11]. This model first generates a "clean" speech reference from the degraded one, then calculates a perceptually-motivated distance between the degraded and the clean speech signal, further extracts a large number of parameters related to typical transmission channel degradations, and combines the perceptually weighted distance and the parameters to an estimation of overall speech quality. Applying this model to synthesized speech [12], the results were not as promising as those obtained with the HMM-based approach, but the parameters have not yet been optimized for synthesized speech. A comparison of different such single-ended speech quality models described in [13] shows that the P.563 model might not be the most appropriate one. In addition, considerable differences have been detected between the performances for male vs. female voices [14].

Our aim is to compare and to combine the feature-comparison and the parametric approaches in order to increase the prediction performance and robustness. For this purpose, we used the optimized HMM-based feature comparison and extracted a number of parameters which correlate with auditory test results, see Section 2. We further collected three auditory databases in order to broaden the basis for a comparison, see Section 3. Applying the approaches to these data-

bases, we analyze the performance and robustness of the predictions in Section 4. Section 5 summarizes the main results and identifies the next steps for further research.

2. Modeling approach

We compare and combine an HMM-based comparison of features with a parametric description of the speech signal in order to derive an estimate of the perceived speech quality. The overall structure is given in Figure 1, and the individual parts are described in the following subsections.

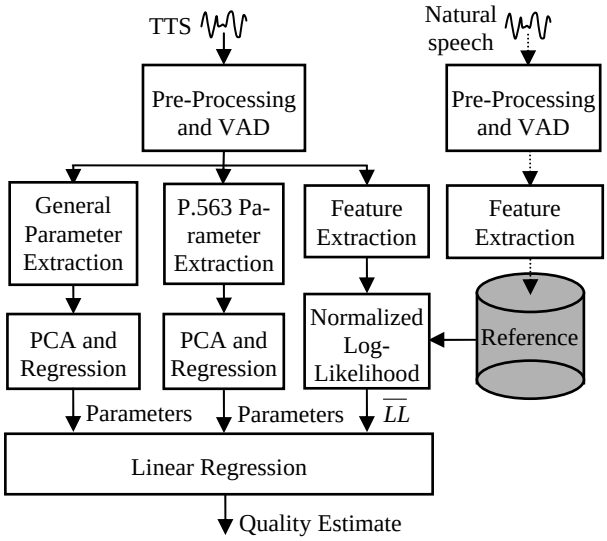

Figure 1: *Modeling approach. Solid lines refer to the evaluation phase, dashed lines to the training phase.*

2.1. HMM-based feature comparison

The HMM-based feature comparison mainly follows the one described in [10]. In order to obtain comparable characteristics for the feature comparison, a pre-processing step is carried out both during the training phase (for the natural speech) and during the evaluation phase (for the TTS samples). It consists of filtering with a standard telephone bandpass (300-3400 Hz) given in ITU-T Rec. G.712, downsampling to 8 kHz, and level normalization to -26 dB below the overload point of the digital system, using the active speech-level meter defined in ITU-T Rec. P.56. Moreover, since we are only interested in the quality of the TTS system, only active speech segments are analyzed, using a simple energy thresholding Voice Activity Detection (VAD) algorithm to remove silence intervals longer than 75 milliseconds; this duration was empirically chosen as to avoid artificial discontinuities introduced by possible VAD errors.

12^{th} order MFCCs are then computed both during the training and the evaluation phase using 25 ms windows and 10 ms time shifts, including the 0^{th} order coefficient which is used as a log-energy measure. In order to quantify signal-energy dynamics, the 0^{th} delta-cepstral coefficient is added which has been shown useful for temporal discontinuity detection. Finally, the fundamental frequency F0 is computed with the pitch tracking algorithm described in [15]. The average F0 over all voiced speech frames is used to identify talker gender, using F0 = 160 Hz as a threshold to distinguish between male and female voices.

Since we consider the temporal dynamics to be important for perceived speech quality, we use HMMs trained with natural reference features to quantify differences between naturally-produced and synthesized speech; HMM transitions should be able to adequately capture the dynamics expressed by the sequence of feature vectors. First experiments show that there is a considerable difference in the features of male and female speech, so we decided to use two reference models, one for male and one for female speech data. HMMs with 8 states are used, the output distribution of each state consisting of a Gaussian mixture density with 16 diagonal-covariance Gaussian components. Model parameters, such as state transition probabilities, initial state probabilities, and output distribution parameters, are computed using the expectation-maximization algorithm [16]. The perceptual similarity is then expressed as a Log-Likelihood (*LL*) value computed using the so-called forward-backward procedure described in [16]. Normalization is performed based on the number of active-speech frames in the signal under test; the normalized log-likelihood is referred to as $\overline{LL}$ in Fig. 1.

2.2. Parameter extraction

As a second basis for the quality estimation, we extracted parameters from the synthesized speech signal which might be related to the degradations coming with the synthesis process. A first set of parameters was taken from the model described in ITU-T Rec. P.563 [11]. These parameters capture characteristics such as noise, temporal clippings, and robotization effects (voice with metallic sounds). A total of 44 characteristic signal parameters are calculated. Based on a restricted set of eight key parameters, one of six major "distortion classes" is detected, such as a high level of background noise, signal interruptions, signal-correlated noise, speech robotization, and unnatural male or female speech. We designate the detected "distortion class" as well as the underlying parameters as the P.563 set of parameters in the following analysis.

Secondly, we calculated a large set of 1495 general parameters [17] which provide a broad variety of information about vocal expression patterns that can be useful when classifying speech metadata such as age, gender and emotion. These parameters are related to signal duration, formants, intensity, loudness, cepstrum, pitch, spectrum, and zero crossing rates. We designate this set as "general parameters" in the following analysis.

In order to extract the relevant information for the given task from this large set of parameters, we employed a sequential feature selection (SFS) algorithm followed by Principal Components Analysis (PCA). The SFS used a correlation-based cost function where features with $|R|>0.25$ were kept. PCA was subsequently used on this subset to come up with a small set of relevant factors which are used for the quality estimation function.

2.3. Linear combination

Finally, a quality estimate is calculated from either $\overline{LL}$, the factors of the principle component analysis of the extracted parameters, or both. Since the available auditory test data is quite limited, we opted for a simple linear regression model which was calculated by the $\overline{LL}$ value and the values given by the linear regression over the PCA factors and estimating the naturalness or overall quality judgment of the particular test. A manual investigation of the shape of the relationship between input variables and auditory judgments did not provide enough evidence for justifying more complicated (e.g. non-linear) relationships.

Input variables	Test 1			Test 2			Test 3		
	R	rmse	ρ	*R*	rmse	ρ	*R*	rmse	ρ
LL	0.77/0.81	0.51/0.45	0.51/0.54	0.48/0.56	0.63/0.54	0.45/0.31	-0.57/-0.98	1.09/0.98	-0.52/-1.00
P.563 params	0.90/0.96	0.34/0.21	0.79/0.83	0.85/0.91	0.37/0.30	0.87/0.89	0.80/0.97	0.37/0.20	0.82/0.50
General params	0.77/0.83	0.57/0.49	0.75/0.71	0.64/0.73	0.57/0.49	0.60/0.71	0.88/1.00	0.30/0.08	0.81/1.00
LL + P.563 params	0.90/0.96	0.36/0.20	0.84/0.83	0.83/0.89	0.38/0.31	0.85/0.94	0.82/0.98	0.36/0.16	0.79/1.00
LL + general params	0.86/0.91	0.43/0.35	0.73/0.66	0.71/0.80	0.47/0.38	0.68/0.60	0.86/0.99	0.32/0.09	0.81/1.00
P.563 + general params	0.89/0.96	0.36/0.22	0.78/0.83	0.82/0.88	0.38/0.31	0.84/0.89	0.76/0.96	0.40/0.23	0.82/0.50
LL + P.563 + general params	0.89/0.96	0.37/0.21	0.81/0.83	0.80/0.85	0.39/0.33	0.82/0.94	0.78/0.98	0.38/0.19	0.77/1.00

Table 1. *Correlations and prediction error for male speech files. Reported values correspond to "per-stimulus/per-synthesizer".*

Input variables	Test 1			Test 2			Test 3		
	R	rmse	ρ	*R*	rmse	ρ	*R*	rmse	ρ
LL	0.17/0.17	0.89/0.86	-0.02/-0.14	-0.32/-0.38	0.80/0.73	-0.35/-0.37	0.38/1.00	0.47/0.18	0.47/1.00
P.563 params	0.81/0.92	0.42/0.31	0.79/0.83	0.48/0.59	0.52/0.43	0.46/0.54	0.69/1.00	0.36/0.22	0.75/1.00
General params	0.89/0.97	0.30/0.16	0.89/0.94	0.80/0.92	0.37/0.25	0.75/0.94	0.69/1.00	0.33/0.16	0.68/1.00
LL + P.563 params	0.84/0.98	0.35/0.16	0.87/0.89	0.53/0.65	0.50/0.39	0.47/0.54	0.73/1.00	0.29/0.12	0.77/1.00
LL + general params	0.89/0.97	0.30/0.15	0.90/0.94	0.81/0.93	0.36/0.23	0.77/0.94	0.70/1.00	0.31/0.14	0.65/1.00
P.563 + general params	0.89/0.97	0.30/0.16	0.89/0.94	0.80/0.92	0.37/0.25	0.75/0.94	0.69/1.00	0.33/0.16	0.68/1.00
LL + P.563 + general params	0.89/0.97	0.29/0.14	0.90/0.94	0.80/0.92	0.36/0.24	0.77/0.94	0.71/1.00	0.31/0.15	0.70/1.00

Table 2. *Correlations and prediction error for female speech files. Reported values correspond to "per-stimulus/per-synthesizer"*

3. Test databases

In order to test our approach, we used data from three auditory tests which have been carried out with different synthesis systems and stimuli, different test participants, and at different places in Germany, for the German language. In the following sub-sections, we briefly describe the experimental set-ups and procedures.

3.1. Test 1

Test 1 has been carried out at the Institute for Phonetics and Digital Speech Processing at Christian-Albrechts-University of Kiel, Germany, see [18] for details. It used speech material from six off-the-shelf TTS systems: 3 commercial ones (AT&T, MBROLA-based Proser, and Cepstral) and three from German academic institutions (TU Dresden, TU Berlin, and University of Bonn), all with male and female voices. A total of 10 speech samples have been generated per TTS system, half for male speakers and half for female ones. The synthesized speech samples have an average duration of 11 seconds and consist of two utterances separated by a silence interval of approximately two seconds. All samples were band-pass-filtered (300-3400 Hz) and normalized to an active speech level of -26 dBov prior to listener presentation.

The listening test procedure closely followed ITU-T Rec. P.85 [3] and was performed in a silent listening room. 17 listeners (10 female, 7 male) participated in the test; all were German students and the age ranged from 20-26. Listeners were given a parallel task and asked to rate the synthesized speech signals using eight different quality rating scales. Of the eight scales used, only five are described in ITU-T Rec. P.85. We limit our analysis here to the ratings on the naturalness (providing the largest range between worst- and best-rated synthesis), but plan to carry out an analysis of the other dimensions in the near future.

3.2. Test 2

The second test was carried out in the frame of a Master thesis at Quality and Usability Lab, TU Berlin. It consisted of speech material from 6 German TTS systems: AT&T, ATIP Proser, DRESS, Nuance RealSpeak, MARY, MBROLA. For each synthesizer, 5 samples of 7-8 s length have been generated, using text material which is typical for train travel announcements. The stimuli have been pre-processed as in Test 1, coded-decoded with log PCM according to ITU-T Rec. G.711, and presented to 25 listeners (13 male, 12 female, age range 20-35years, average age 25.8) in a quiet test room. Again, the procedure described in ITU-T Rec. P.85 was followed, using 4 rating scales of which only the naturalness scale will be further analyzed here.

3.3. Test 3

The third test was part of a Bachelor thesis carried out at Ruhr-University Bochum and is described in detail in [19]. It contained speech stimuli from 3 different TTS systems (SyRUB, the MBROLA-based Proser and the Cepstral synthesizer, the latter two with two different voices each, resulting in only 3 male and 2 female voices), as well as two natural voices. The speech samples have been transmitted over a simulated standard ISDN telephone channel with default characteristics, and then judged for their overall quality in a natural living-room environment which did not fully respect the acoustic requirements for test rooms given in ITU-T Rec. P.800. In contrast to Tests 1 and 2, no parallel task was given to the test subjects; they just had to rate the overall quality (not "naturalness") on a continuous scale labeled from 0 to 6, using a slider presented on a computer screen. 20 naïve listeners (10 male, 10 female, no age record available) participated in the test, most of them were university students.

4. Results and discussion

The subjective ratings have been averaged per stimulus which can be compared to the estimated quality rating obtained from the model. We used the Log-Likelihood, the P.563 parameters, the general parameters, and any combination of these as input parameters to the quality estimation function, and report on the correlations and the root mean squared error. Since the rating scale has not really interval level we provide both the Pearson correlation *R* and the Spearman rank-order correlation ρ. The analysis is first carried out on a per-stimulus basis and

then on a per-synthesizer basis; it is limited to the synthesized speech samples only, as we did not want to artificially increase the correlations by adding the naturally-produced stimuli which usually show a higher quality and thus increase the range of quality levels covered in the experiment.

The results for the male stimuli are shown in Tab. 1. Whereas all input variables work relatively well for predicting the Test 1 data, *LL* shows problems with the data from Test 2 and Test 3. For these tests, the parameter-based approach is significantly better, both with the P.563 parameters and with the general parameter set. The best combination which shows $R > 0.8$ for all databases is *LL* and the P.563 parameters.

For the female stimuli, only the general parameters show a satisfying performance; here, the combination of *LL* and general parameters reaches $R > 0.8$ for Tests 1 and 2, and still $R = 0.7$ for Test 3. It is important to note that the low correlation in Test 1 does not contradict the higher performance which has been observed in [10], as one of the synthesizers (MBROLA, TU Berlin) was excluded from that analysis, providing a significantly higher correlation. Up to now, we cannot explain that outlier, but we kept it in the analysis to highlight the problems of the *LL* approach with female data.

On a per-synthesizer basis, the performance of the estimators still increases. Apparently, the differences between individually synthesized speech samples are averaged out in the per-synthesizer analysis. This shows that the best-performing models stated above are even better in comparing different synthesizers than comparing individual synthesized speech samples. However, it has to be noted that Test 3 only contained 3 male and 2 female synthesizers; the high correlations get meaningless in that case.

5. Conclusions and future work

We compared and combined two approaches for instrumentally predicting TTS quality on 3 auditory databases. Over all databases, the combination of *LL* with the P.563 parameters achieved the best performance on the male data, and the combination of *LL* and the general parameters on the female data. Lower performance was obtained using the *LL* approach on female data relative to male data. While the source of this gap is still unknown, we suspect it may be due to errors in online F0 calculation, thus leading female speech signals to be scored against male reference HMMs.

Overall, the correlations obtained on all databases are quite encouraging: In 5 out of 6 cases, correlations greater than 0.8 could be obtained; further increases could be observed on a per-synthesizer basis. As we expected, this indicates that the approach is better for differentiating between synthesizers than it is for differentiating between individual stimuli produced by one particular synthesizer.

We plan to extend the analysis to the other rating scales of the auditory experiments. We will try to find underlying reasons for the bad performance of the *LL* approach on the female data, and further analyze the impact of the natural speech data used for training the reference models. We would like to come up with one model fitting both male and female synthesized voices, and different types of (also formant) synthesizers, and test it on independent data (e.g. from the Blizzard challenge) in order to analyze the robustness of our approach.

6. Acknowledgement

The authors would like to thank Kathrin Seget and Ulrich Heute from Christian-Albrechts-University of Kiel for making available the Test 1 data, and Johannes Heimansberg (formerly IKA, Ruhr-Universität Bochum) for the Test 3 data. The work is partly supported by the Deutsche Forschungsgemeinschaft, DFG (project MO 1038/11-1).

7. References

[1] Van Bezooijen, R. and van Heuven, V.J., "Assessment of Speech Output Systems", in: D. Gibbon, R. Moore and R. Winski [Eds.], Handbook of Standards and Resources for Spoken Language Systems, 481-563, Mouton de Gruyter, Berlin, 1997.

[2] Delogu, C., Conte, S. and Sementina, C., "Cognitive Factors in the Evaluation of Synthetic Speech", Speech Communication, 24(2):153-168, 1998.

[3] ITU-T Rec. P.85, "A Method for Subjective Performance Assessment of the Quality of Speech Voice Output Devices", Int. Telecomm. Union, Geneva, 1994.

[4] Sityaev, D., Knill, K. and Burrows, T., "Comparison of the ITU-T P.85 Standard to Other Methods for the Evaluation of Text-to-Speech Systems", in: Proc. 9th Int. Conf. on Spoken Language Process. (Interspeech 2006 – ICSLP), Pittsburgh PA, 1077-1080, 2006.

[5] Viswanathan, M. and 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 19:55-83, 2005.

[6] Cernak, M. and Rusko, M., "An Evaluation of Synthetic Speech Using the PESQ Measure", in: Proc. European Congress on Acoustics, 2725-2728, 2005.

[7] Quackenbush, S.R., Barnwell, T.P. and Clements, M.A., "Objective Measures of Speech Quality", Prentice Hall, Englewood Cliffs, 1988.

[8] ITU-T Rec. P.862, "Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs", Int. Telecomm. Union, Geneva, 2001.

[9] Mariniak, A., "A Global Framework for the Assessment of Synthetic Speech Without Subjects", in: Proc. 3rd Europ. Conf. on Speech Process. And Technology (Eurospeech'93), Berlin, 1683-1686, 1993.

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

[11] ITU-T Rec. P.563, "Single-ended Method for Objective Speech Quality Assessment in Narrow-band Telephony Applications", Int. Telecomm. Union, Geneva, 2004.

[12] Falk, T.H., Möller, S., Karaiskos, V. and King, S., "Improving Instrumental Quality Prediction Performance for the Blizzard Challenge", in: Proc. Blizzard Challenge Workshop, Brisbane, 6 pages, 2008.

[13] Möller, S., Kim, D.-S. and Malfait, L., "Estimating the Quality of Synthesized and Natural Speech Transmitted Through Telephone Networks Using Single-ended Prediction Models", Acta Acustica united with Acustica 94:21-31, 2008.

[14] ITU-T Contr. COM 12-180, "Single-Ended Quality Estimation of Synthesized Speech: Analysis of the Rec. P.563 Internal Signal Processing", Source: Federal Republic of Germany (Authors: S. Möller, T.H. Falk), ITU-T SG12 Meeting, 22-29 May 2008, Geneva.

[15] Talkin, D., "A Robust Algorithm for Pitch Tracking (RAPT)", in: Kleijn, W.B. and Paliwal, K.K., eds., Speech Coding and Synthesis, Elsevier Science Publishers, Amsterdam, 495–518, 1995.

[16] Rabiner, L., "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," Proceedings of the IEEE 77(2): 257–286, Feb. 1989.

[17] Minker, W. Lee, G.G., Mariani, J. and Nakamura, S., Spoken Dialogue Systems Technology and Design, Springer, Boston MA, 2010.

[18] Seget, K., "Untersuchungen zur auditiven Qualität von Sprachsyntheseverfahren" (Study of perceptual quality of text-to-speech systems), July 2007, Diploma thesis, Christian-Albrechts-University of Kiel.

[19] Möller, S. and Heimansberg, J., "Estimation of TTS Quality in Telephone Environments Using a Reference-free Quality Prediction Model", in: Proc. 2nd ISCA/DEGA Tutorial and Research Workshop on Perceptual Quality of Systems, Berlin, 56-60, 2006.